10/649,100

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

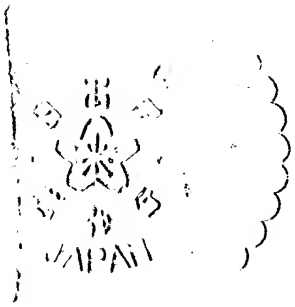
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年　１月３１日

出願番号
Application Number: 特願２００３－０２５０７４

[ST. 10/C]: [ＪＰ２００３－０２５０７４]

出願人
Applicant(s): 株式会社日立製作所



２００３年　８月１３日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００３－３０６５１６１

【書類名】　特許願
【整理番号】　HI020901
【提出日】　平成15年 1月31日
【あて先】　特許庁長官殿
【国際特許分類】　G06F　3/06
【発明者】
　【住所又は居所】　神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所ソフトウェア事業部内
　【氏名】　藤本　修二
【特許出願人】
　【識別番号】　000005108
　【氏名又は名称】　株式会社日立製作所
【代理人】
　【識別番号】　100071283
　【弁理士】
　【氏名又は名称】　一色　健輔
【選任した代理人】
　【識別番号】　100084906
　【弁理士】
　【氏名又は名称】　原島　典孝
【選任した代理人】
　【識別番号】　100098523
　【弁理士】
　【氏名又は名称】　黒川　恵
【選任した代理人】
　【識別番号】　100112748
　【弁理士】
　【氏名又は名称】　吉田　浩二

【選任した代理人】

【識別番号】　100110009

【弁理士】

【氏名又は名称】　青木　康

【手数料の表示】

【予納台帳番号】　011785

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　記憶デバイス制御装置、及びプログラム

【特許請求の範囲】

【請求項１】　情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応するＩ／Ｏ要求を出力するＩ／Ｏプロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、

前記各チャネル制御部はフェイルオーバを行うためのグループに類別されている記憶デバイス制御装置であって、

前記各チャネル制御部により更新される前記フェイルオーバ時に引き継がれるデータを、当該チャネル制御部と同一の前記グループに類別されている前記各チャネル制御部が共通にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である共有ボリュームに記憶する手段を備えることを特徴とする記憶デバイス制御装置。

【請求項２】　情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応するＩ／Ｏ要求を出力するＩ／Ｏプロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、

前記各チャネル制御部はフェイルオーバを行うためのグループに類別されている記憶デバイス制御装置であって、

前記各チャネル制御部により更新される前記フェイルオーバ時に引き継がれるデータを、前記各チャネル制御部が共通にアクセス可能な前記記憶デバイス制御装置が備える共有メモリに記憶する手段を備えることを特徴とする記憶デバイス制御装置。

【請求項３】　情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応するＩ／Ｏ要求を出力するＩ／Ｏプロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、

前記各チャネル制御部はフェイルオーバを行うためのグループに類別されてい

る記憶デバイス制御装置であって、

前記各チャネル制御部により更新される前記フェイルオーバ時に引き継がれるデータを、当該チャネル制御部と同一の前記グループに類別されている他の前記チャネル制御部に、前記各チャネル制御部を相互に接続するネットワークを介して送信する手段を備えることを特徴とする記憶デバイス制御装置。

【請求項４】　請求項１又は請求項２に記載の記憶デバイス制御装置において、

前記各チャネル制御部には、当該チャネル制御部が個別にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域であるローカルボリュームがそれぞれ割り当てられており、

前記データを更新した前記チャネル制御部と同一の前記グループに類別されている他の前記チャネル制御部の前記ローカルボリュームにも前記データを記憶する手段をさらに備えることを特徴とする記憶デバイス制御装置。

【請求項５】　請求項１又は請求項２に記載の記憶デバイス制御装置において、

前記各チャネル制御部には、当該チャネル制御部が個別にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域であるローカルボリュームがそれぞれ割り当てられており、

前記データを更新した前記チャネル制御部と同一の前記グループに類別されている他の前記チャネル制御部の前記ローカルボリュームにも前記データを記憶する手段と、

前記データの参照先が記録された引き継ぎデータ参照テーブルと、

前記引き継ぎデータ参照テーブルに記録された前記データの参照先に基づき、前記共有ボリューム、前記共有メモリ、又は前記ローカルボリュームのいずれかから前記データを読み出す手段と

をさらに備えることを特徴とする記憶デバイス制御装置。

【請求項６】　請求項３に記載の記憶デバイス制御装置において、

前記データが前記記憶デバイス制御装置内の全ての前記チャネル制御部に共通に参照されるデータである場合には、当該データを、前記ネットワークを介して

、前記記憶デバイス制御装置内の全ての前記チャネル制御部に送信する手段を備えることを特徴とする記憶デバイス制御装置。

【請求項７】　請求項１に記載の記憶デバイス制御装置において、

前記データが前記記憶デバイス制御装置内の全ての前記チャネル制御部に共通に参照されるデータである場合には、当該データを、前記記憶デバイス制御装置内の全ての前記チャネル制御部が共通にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である第２の共有ボリュームに記憶する手段を備えることを特徴とする記憶デバイス制御装置。

【請求項８】　請求項１、請求項２、又は請求項３に記載の記憶デバイス制御装置において、

前記フェイルオーバ時に引き継がれるデータには、少なくとも

ＮＦＳユーザデータ、ＣＩＦＳユーザデータ、装置管理者データ、フェイルオーバ用ハートビート、チャネル制御部のＩＰアドレス、ＮＦＳファイルロック情報、クラスタ制御情報のいずれかが含まれることを特徴とする記憶デバイス制御装置。

【請求項９】　情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応するＩ／Ｏ要求を出力するＩ／Ｏプロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、

前記各チャネル制御部はフェイルオーバを行うためのグループに類別されている記憶デバイス制御装置に、

前記各チャネル制御部により更新される前記フェイルオーバ時に引き継がれるデータを、当該チャネル制御部と同一の前記グループに類別されている前記各チャネル制御部が共通にアクセス可能な前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である共有ボリュームに記憶するステップを実行させるためのプログラム。

【請求項１０】　情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応するＩ／Ｏ要求を出力するＩ／Ｏプロセッサとが形成された回

路基板を有する複数のチャネル制御部を備え、

前記各チャネル制御部はフェイルオーバを行うためのグループに類別されている記憶デバイス制御装置に、

前記各チャネル制御部により更新される前記フェイルオーバ時に引き継がれるデータを、前記各チャネル制御部が共通にアクセス可能な前記記憶デバイス制御装置が備える共有メモリに記憶するステップを実行させるためのプログラム。

【請求項１１】　情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応するＩ／Ｏ要求を出力するＩ／Ｏプロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、

前記各チャネル制御部はフェイルオーバを行うためのグループに類別されている記憶デバイス制御装置に、

前記各チャネル制御部により更新される前記フェイルオーバ時に引き継がれるデータを、当該チャネル制御部と同一の前記グループに類別されている他の前記チャネル制御部に、前記各チャネル制御部を相互に接続するネットワークを介して送信するステップを実行させるためのプログラム。

【請求項１２】　請求項９又は請求項１０に記載のプログラムにおいて、

前記各チャネル制御部には、当該チャネル制御部が個別にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域であるローカルボリュームがそれぞれ割り当てられており、

前記データを更新した前記チャネル制御部と同一の前記グループに類別されている他の前記チャネル制御部の前記ローカルボリュームにも前記データを記憶するステップをさらに実行させるためのプログラム。

【請求項１３】　請求項９又は請求項１０に記載のプログラムにおいて、

前記各チャネル制御部には、当該チャネル制御部が個別にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域であるローカルボリュームがそれぞれ割り当てられており、

前記データを更新した前記チャネル制御部と同一の前記グループに類別されている他の前記チャネル制御部の前記ローカルボリュームにも前記データを記憶す

るステップと、

前記データの参照先が記録された引き継ぎデータ参照テーブルを参照するステップと、

前記引き継ぎデータ参照テーブルに記録された前記データの参照先に基づき、前記共有ボリューム、前記共有メモリ、又は前記ローカルボリュームのいずれかから前記データを読み出すステップと

をさらに実行させるためのプログラム。

【請求項１４】　請求項１１に記載のプログラムにおいて、

前記データが前記記憶デバイス制御装置内の全ての前記チャネル制御部に共通に参照されるデータである場合には、当該データを、前記ネットワークを介して、前記記憶デバイス制御装置内の全ての前記チャネル制御部に送信するステップを前記記憶デバイス制御装置に実行させるためのプログラム。

【請求項１５】　請求項９に記載のプログラムにおいて、

前記データが前記記憶デバイス制御装置内の全ての前記チャネル制御部に共通に参照されるデータである場合には、当該データを、前記記憶デバイス制御装置内の全ての前記チャネル制御部が共通にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である第２の共有ボリュームに記憶するステップを前記記憶デバイス制御装置に実行させるためのプログラム。

【請求項１６】　請求項９、請求項１０、又は請求項１１に記載のプログラムにおいて、

前記フェイルオーバ時に引き継がれるデータには、少なくとも

ＮＦＳユーザデータ、ＣＩＦＳユーザデータ、装置管理者データ、フェイルオーバ用ハートビート、チャネル制御部のＩＰアドレス、ＮＦＳファイルロック情報、クラスタ制御情報のいずれかが含まれることを特徴とするプログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、記憶デバイス制御装置、及びプログラムに関する。

【０００２】

【従来の技術】

近年コンピュータシステムで取り扱われるデータ量が急激に増加している。このようなデータを管理するためのストレージシステムとして、最近ではミッドレンジクラスやエンタープライズクラスと呼ばれるような、巨大な記憶資源を提供するＲＡＩＤ（Redundant Arrays of Inexpensive Disks）方式で管理された大規模なストレージシステムが注目されている。

【０００３】

また、ストレージシステムと情報処理装置とをＴＣＰ／ＩＰ（Transmission Control Protocol/Internet Protocol）プロトコル等を用いたネットワークで相互に接続し、情報処理装置からのファイルレベルでのアクセスを実現するＮＡＳ（Network Attached Storage）と呼ばれるストレージシステムが開発されている。

【０００４】

一方、ストレージシステムが提供する記憶領域を使用する情報処理装置に障害が発生しても、他の情報処理装置に処理を代替させることにより、情報処理装置によるサービスの提供を継続させることができるフェイルオーバと呼ばれる技術が開発されている。

【０００５】

【特許文献１】

特開平８－２１２０９５号公報

【０００６】

【発明が解決しようとする課題】

しかしながら従来のストレージシステムにおけるフェイルオーバでは、情報処理装置間を跨る共通の記憶領域が設けられていないため、フェイルオーバを行う場合には、元の情報処理装置が使用していた記憶領域のデータを、処理を代替する情報処理装置が使用できるように引き継ぎ処理をしなければならない。また情報処理装置に障害が発生した後に記憶領域のデータの引き継ぎが行われるため、代替処理が開始されるまでにタイムラグが発生していた。さらにデータの引き継

ぎのための処理が煩雑であった。

本発明は上記課題を鑑みてなされたものであり、記憶デバイス制御装置、及びプログラムを提供することを主たる目的とする。

【０００７】

【課題を解決するための手段】

上記課題を解決するために、本発明に係る記憶デバイス制御装置は、情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応するＩ／Ｏ要求を出力するＩ／Ｏプロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、前記各チャネル制御部はフェイルオーバを行うためのグループに類別されている記憶デバイス制御装置であって、前記各チャネル制御部により更新される前記フェイルオーバ時に引き継がれるデータを、当該チャネル制御部と同一の前記グループに類別されている前記各チャネル制御部が共通にアクセス可能な前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である共有ボリュームに記憶する手段を備える。

【０００８】

なお、前記情報処理装置とは、前記構成の前記記憶デバイス制御装置を備えて構成されるストレージシステムにＬＡＮ（Local Area Network）を介してアクセスする、例えば、パーソナルコンピュータやメインフレームコンピュータである。ファイルアクセス処理部の機能はＣＰＵ上で実行されるオペレーティングシステムおよびこのオペレーティングシステム上で動作する例えばＮＦＳ（Network File System）等のソフトウエアによって提供される。記憶デバイスは例えばハードディスク装置などのディスクドライブである。Ｉ／Ｏプロセッサは例えばファイルアクセス処理部のハードウエア要素である前記ＣＰＵとは独立したＩＣ（Integrated Circuit）をハードウエア要素とし、ファイルアクセス処理部とディスク制御部との間の通信を制御する。ディスク制御部は、記憶デバイスに対してデータの書き込みや読み出しを行う。

その他、本願が開示する課題、及びその解決方法は、発明の実施の形態の欄、及び図面により明らかにされる。

【０００９】

【発明の実施の形態】

以下、本発明の実施の形態について図面を用いて詳細に説明する。

まず、本実施の形態に係るストレージシステム６００の全体構成を示すブロック図を図１に示す。

＝＝＝全体構成例＝＝＝

ストレージシステム６００は、記憶デバイス制御装置１００と記憶デバイス３００とを備えている。記憶デバイス制御装置１００は、情報処理装置２００から受信したコマンドに従って記憶デバイス３００に対する制御を行う。例えば情報処理装置２００からデータの入出力要求を受信して、記憶デバイス３００に記憶されているデータの入出力のための処理を行う。データは、記憶デバイス３００が備えるディスクドライブにより提供される物理的な記憶領域上に論理的に設定される記憶領域である論理ボリューム（Logical Unit）（以下、ＬＵと記す）に記憶されている。また記憶デバイス制御装置１００は、情報処理装置２００との間で、ストレージシステム６００を管理するための各種コマンドの授受も行う。

【００１０】

情報処理装置２００はＣＰＵ（Central Processing Unit）やメモリを備えたコンピュータである。情報処理装置２００が備えるＣＰＵにより各種プログラムが実行されることにより様々な機能が実現される。情報処理装置２００は、例えばパーソナルコンピュータやワークステーションであることもあるし、メインフレームコンピュータであることもある。

【００１１】

図１において、情報処理装置２００は、ＬＡＮ（Local Area Network）４００を介して記憶デバイス制御装置１００と接続されている。ＬＡＮ４００は、インターネットとすることもできるし、専用のネットワークとすることもできる。ＬＡＮ４００を介して行われる情報処理装置２００と記憶デバイス制御装置１００との間の通信は、例えばＴＣＰ／ＩＰプロトコルに従って行われる。情報処理装置２００からは、ストレージシステム６００に対して、ファイル名指定によるデータアクセス要求（ファイル単位でのデータ入出力要求。以下、ファイルアクセ

ス要求と記す）が送信される。

【００１２】

記憶デバイス制御装置１００は、チャネル制御部１１０を備える。以下、このチャネル制御部１１０のことをＣＨＮとも記す。記憶デバイス制御装置１００は、チャネル制御部１１０によりＬＡＮ４００を介して情報処理装置２００との間で通信を行う。チャネル制御部１１０は、情報処理装置２００からのファイルアクセス要求を個々に受け付ける。すなわち、チャネル制御部１１０には、個々にＬＡＮ４００上のネットワークアドレス（例えば、ＩＰアドレス）が割り当てられていてそれぞれが個別にＮＡＳとして振る舞い、個々のＮＡＳがあたかも独立したＮＡＳが存在しているかのようにＮＡＳのサービスを情報処理装置２００に提供することができる。以下、チャネル制御部１１０をＣＨＮと記す。このように１台のストレージシステム６００に個別にＮＡＳとしてのサービスを提供するチャネル制御部１１０を備えるように構成したことで、従来、独立したコンピュータで個別に運用されていたＮＡＳサーバが一台のストレージシステム６００に集約されて運用される。そして、この構成によってストレージシステム６００の統括的な管理が可能となり、各種設定・制御や障害管理、バージョン管理といった保守業務の効率化が図られる。

【００１３】

なお、本実施の形態に係る記憶デバイス制御装置１００のチャネル制御部１１０の機能は、後述するように、一体的にユニット化された回路基板上に形成されたハードウエア及びこのハードウエアにより実行されるオペレーティングシステム（以下、ＯＳと記す）やこのＯＳ上で動作するアプリケーションプログラムなどのソフトウエアにより実現される。このように本実施例のストレージシステム６００では、従来ハードウエアの一部として実装されてきた機能が主としてソフトウエアによって実現されている。このため、本実施例のストレージシステム６００では柔軟性に富んだシステム運用が可能となり、多様で変化の激しいユーザニーズに対応したきめ細かなサービスを提供することが可能となる。

【００１４】

＝＝＝記憶デバイス＝＝＝

記憶デバイス３００は多数のディスクドライブ（物理ディスク）を備えており情報処理装置２００に対して記憶領域を提供する。データは、ディスクドライブにより提供される物理的な記憶領域上に論理的に設定される記憶領域であるＬＵに記憶されている。ディスクドライブとしては、例えばハードディスク装置やフレキシブルディスク装置、半導体記憶装置等様々なものを用いることができる。

【００１５】

なお、記憶デバイス３００は例えば複数のディスクドライブによりディスクアレイを構成するようにすることもできる。この場合、情報処理装置２００に対して提供される記憶領域は、ＲＡＩＤにより管理された複数のディスクドライブにより提供されるようにすることもできる。

記憶デバイス制御装置１００と記憶デバイス３００との間は図１のように直接に接続される形態とすることもできるし、ネットワークを介して接続するようにすることもできる。さらに記憶デバイス３００は記憶デバイス制御装置１００と一体として構成されることもできる。

【００１６】

記憶デバイス３００に設定されるＬＵには、情報処理装置２００からアクセス可能なユーザＬＵや、チャネル制御部１１０の制御のために使用されるシステムＬＵ等がある。システムＬＵにはＣＨＮ１１０で実行されるオペレーティングシステムも格納される。また各ＬＵにはチャネル制御部１１０が対応付けられている。これによりチャネル制御部１１０毎にアクセス可能なＬＵが割り当てられている。また上記対応付けは、複数のチャネル制御部１１０で一つのＬＵを共有するようにすることもできる。なお以下において、ユーザＬＵやシステムＬＵをユーザディスク、システムディスク等とも記す。また、複数のチャネル制御部１１０で共有されるＬＵを共有ＬＵあるいは共有ディスクと記す。

【００１７】

＝＝＝記憶デバイス制御装置＝＝＝

記憶デバイス制御装置１００はチャネル制御部１１０、共有メモリ１２０、キャッシュメモリ１３０、ディスク制御部１４０、管理端末１６０、接続部１５０を備える。

【００１８】

チャネル制御部１１０は情報処理装置２００との間で通信を行うための通信インタフェースを備え、情報処理装置２００との間でデータ入出力コマンド等を授受する機能を備える。例えばＣＨＮ１１０は情報処理装置１乃至３（２００）からのファイルアクセス要求を受け付ける。そしてファイルの記憶アドレスやデータ長等を求めて、ファイルアクセス要求に対応するＩ／Ｏ要求を出力することにより、記憶デバイス３００へのアクセスを行う。これによりストレージシステム６００はＮＡＳとしてのサービスを情報処理装置１乃至３（２００）に提供することができる。なおＩ／Ｏ要求にはデータの先頭アドレス、データ長、読み出し又は書き込み等のアクセスの種別が含まれている。またデータの書き込みの場合にはＩ／Ｏ要求には書き込みデータが含まれているようにすることもできる。Ｉ／Ｏ要求の出力は、後述するＩ／Ｏプロセッサ１１９により行われる。

【００１９】

各チャネル制御部１１０は管理端末１６０と共に内部ＬＡＮ１５１で接続されている。これによりチャネル制御部１１０に実行させるマイクロプログラム等を管理端末１６０から送信しインストールすることが可能となっている。チャネル制御部１１０の構成については後述する。

【００２０】

接続部１５０はチャネル制御部１１０、共有メモリ１２０、キャッシュメモリ１３０、ディスク制御部１４０を相互に接続する。チャネル制御部１１０、共有メモリ１２０、キャッシュメモリ１３０、ディスク制御部１４０間でのデータやコマンドの授受は接続部１５０を介することにより行われる。接続部１５０は例えば高速スイッチングによりデータ伝送を行う超高速クロスバスイッチなどの高速バスである。チャネル制御部１１０同士が高速バスで接続されていることで、個々のコンピュータ上で動作するＮＡＳサーバがＬＡＮを通じて接続する従来の構成に比べてチャネル制御部１１０間の通信パフォーマンスが大幅に向上する。またこれにより高速なファイル共有機能や高速フェイルオーバなども可能となる。

【００２１】

共有メモリ１２０及びキャッシュメモリ１３０は、チャネル制御部１１０、ディスク制御部１４０により共有される記憶メモリである。共有メモリ１２０は主に制御情報やコマンド等を記憶するために利用されるのに対し、キャッシュメモリ１３０は主にデータを記憶するために利用される。

【００２２】

例えば、あるチャネル制御部１１０が情報処理装置２００から受信したデータ入出力コマンドが書き込みコマンドであった場合には、当該チャネル制御部１１０は書き込みコマンドを共有メモリ１２０に書き込むと共に、情報処理装置２００から受信した書き込みデータをキャッシュメモリ１３０に書き込む。一方、ディスク制御部１４０は共有メモリ１２０を監視しており、共有メモリ１２０に書き込みコマンドが書き込まれたことを検出すると、当該コマンドに従ってキャッシュメモリ１３０から書き込みデータを読み出して記憶デバイス３００に書き込む。

【００２３】

またあるチャネル制御部１１０が情報処理装置２００から受信したデータ入出力コマンドが読み出しコマンドであった場合には、当該チャネル制御部１１０は読み出しコマンドを共有メモリ１２０に書き込むと共に、読み出し対象となるデータがキャッシュメモリ１３０に存在するかどうかを調べる。ここでキャッシュメモリ１３０に存在すれば、チャネル制御部１１０はそのデータを情報処理装置２００に送信する。一方、読みだし対象となるデータがキャッシュメモリ１３０に存在しない場合には、共有メモリ１２０を監視することにより読み出しコマンドが共有メモリ１２０に書き込まれたことを検出したディスク制御部１４０が、記憶デバイス３００から読みだし対象となるデータを読み出してこれをキャッシュメモリ１３０に書き込むと共に、その旨を共有メモリ１２０に書き込む。そして、チャネル制御部１１０は共有メモリ１２０を監視することにより読みだし対象となるデータがキャッシュメモリ１３０に書き込まれたことを検出すると、そのデータを情報処理装置２００に送信する。

【００２４】

なお、このようにチャネル制御部１１０からディスク制御部１４０に対するデ

ータの書き込みや読み出しの指示を共有メモリ１２０を介在させて間接に行う構成の他、例えばチャネル制御部１１０からディスク制御部１４０に対してデータの書き込みや読み出しの指示を共有メモリ１２０を介さずに直接に行う構成とすることもできる。

【００２５】

ディスク制御部１４０は記憶デバイス３００の制御を行う。例えば上述のように、チャネル制御部１１０が情報処理装置２００から受信したデータ書き込みコマンドに従って記憶デバイス３００へデータの書き込みを行う。また、チャネル制御部１１０により送信された論理アドレス指定によるＬＵへのデータアクセス要求を、物理アドレス指定による物理ディスクへのデータアクセス要求に変換する。記憶デバイス３００における物理ディスクがＲＡＩＤにより管理されている場合には、ＲＡＩＤ構成（例えば、ＲＡＩＤ０，１，５）に従ったデータのアクセスを行う。またディスク制御部１４０は、記憶デバイス３００に記憶されたデータの複製管理の制御やバックアップ制御を行う。さらにディスク制御部１４０は、災害発生時のデータ消失防止（ディザスタリカバリ）などを目的としてプライマリサイトのストレージシステム６００のデータの複製をセカンダリサイトに設置された他のストレージシステムにも記憶する制御（レプリケーション機能、又はリモートコピー機能）なども行う。

【００２６】

各ディスク制御部１４０は管理端末１６０と共に内部ＬＡＮ１５１で接続されており、相互に通信を行うことが可能である。これにより、ディスク制御部１４０に実行させるマイクロプログラム等を管理端末１６０から送信しインストールすることが可能となっている。ディスク制御部１４０の構成については後述する。

【００２７】

本実施例においては、共有メモリ１２０及びキャッシュメモリ１３０がチャネル制御部１１０及びディスク制御部１４０に対して独立に設けられていることについて記載したが、本実施例はこの場合に限られるものでなく、共有メモリ１２０又はキャッシュメモリ１３０がチャネル制御部１１０及びディスク制御部１４

０の各々に分散されて設けられることも好ましい。この場合、接続部１５０は、分散された共有メモリ又はキャッシュメモリを有するチャネル制御部１１０及びディスク制御部１４０を相互に接続させることになる。

【００２８】

＝＝＝管理端末＝＝＝

管理端末１６０はストレージシステム６００を保守・管理するためのコンピュータである。管理端末１６０を操作することにより、例えば記憶デバイス３００内の物理ディスク構成の設定や、ＬＵの設定、チャネル制御部１１０において実行されるマイクロプログラムのインストール等を行うことができる。ここで、記憶デバイス３００内の物理ディスク構成の設定としては、例えば物理ディスクの増設や減設、ＲＡＩＤ構成の変更（例えばＲＡＩＤ１からＲＡＩＤ５への変更等）等を行うことができる。さらに管理端末１６０からは、ストレージシステム６００の動作状態の確認や故障部位の特定、チャネル制御部１１０で実行されるオペレーティングシステムのインストール等の作業を行うこともできる。また管理端末１６０はＬＡＮや電話回線等で外部保守センタと接続されており、管理端末１６０を利用してストレージシステム６００の障害監視を行ったり、障害が発生した場合に迅速に対応することも可能である。障害の発生は例えばＯＳやアプリケーションプログラム、ドライバソフトウエアなどから通知される。この通知はＨＴＴＰプロトコルやＳＮＭＰ（Simple Network Management Protocol）、電子メールなどにより行われる。これらの設定や制御は、管理端末１６０で動作するＷｅｂサーバが提供するＷｅｂページをユーザインタフェースとしてオペレータなどにより行われる。オペレータ等は、管理端末１６０を操作して障害監視する対象や内容の設定、障害通知先の設定などを行うこともできる。

【００２９】

管理端末１６０は記憶デバイス制御装置１００に内蔵されている形態とすることもできるし、外付けされている形態とすることもできる。また管理端末１６０は、記憶デバイス制御装置１００及び記憶デバイス３００の保守・管理を専用に行うコンピュータとすることもできるし、汎用のコンピュータに保守・管理機能を持たせたものとすることもできる。

【００３０】

管理端末１６０の構成を示すブロック図を図２に示す。

管理端末１６０は、ＣＰＵ１６１、メモリ１６２、ポート１６３、記録媒体読取装置１６４、入力装置１６５、出力装置１６６、記憶装置１６８を備える。

【００３１】

ＣＰＵ１６１は管理端末１６０の全体の制御を司るもので、メモリ１６２に格納されたプログラム１６２ｃを実行することにより上記Ｗｅｂサーバとしての機能等を実現する。メモリ１６２には、物理ディスク管理テーブル１６２ａとＬＵ管理テーブル１６２ｂとプログラム１６２ｃとが記憶されている。

【００３２】

物理ディスク管理テーブル１６２ａは、記憶デバイス３００に備えられる物理ディスク（ディスクドライブ）を管理するためのテーブルである。物理ディスク管理テーブル１６２ａを図３に示す。図３においては、記憶デバイス３００が備える多数の物理ディスクのうち、ディスク番号＃００１乃至＃００６までが示されている。それぞれの物理ディスクに対して、容量、ＲＡＩＤ構成、使用状況が示されている。

【００３３】

ＬＵ管理テーブル１６２ｂは、上記物理ディスク上に論理的に設定されるＬＵを管理するためのテーブルである。ＬＵ管理テーブル１６２ｂを図４に示す。図４においては、記憶デバイス３００上に設定される多数のＬＵのうち、ＬＵ番号＃ｉ乃至＃３までが示されている。それぞれのＬＵに対して、物理ディスク番号、容量、ＲＡＩＤ構成が示されている。

【００３４】

記録媒体読取装置１６４は、記録媒体１６７に記録されているプログラムやデータを読み取るための装置である。読み取られたプログラムやデータはメモリ１６２や記憶装置１６８に格納される。従って、例えば記録媒体１６７に記録されたプログラム１６２ｃを、記録媒体読取装置１６４を用いて上記記録媒体１６７から読み取って、メモリ１６２や記憶装置１６８に格納するようにすることができる。記録媒体１６７としてはフレキシブルディスクやＣＤ－ＲＯＭ、ＤＶＤ－

ＲＯＭ、ＤＶＤ－ＲＡＭ、半導体メモリ等を用いることができる。なお、上記プログラム１６２ｃは管理端末１６０を動作させるためのプログラムとすることができる他、チャネル制御部１１０やディスク制御部１４０にＯＳ７０１やアプリケーションプログラムをインストールするためのプログラムや、バージョンアップするためのプログラムとすることもできる。記録媒体読取装置１６４は管理端末１６０に内蔵されている形態とすることもできるし、外付されている形態とすることもできる。記憶装置１６８は、例えばハードディスク装置やフレキシブルディスク装置、半導体記憶装置等である。入力装置１６５はオペレータ等による管理端末１６０へのデータ入力等のために用いられる。入力装置１６５としては例えばキーボードやマウス等が用いられる。出力装置１６６は情報を外部に出力するための装置である。出力装置１６６としては例えばディスプレイやプリンタ等が用いられる。ポート１６３は内部ＬＡＮ１５１に接続されており、これにより管理端末１６０はチャネル制御部１１０やディスク制御部１４０等と通信を行うことができる。またポート１６３は、ＬＡＮ４００に接続するようにすることもできるし、電話回線に接続するようにすることもできる。

【００３５】

＝＝＝外観図＝＝＝

次に、本実施の形態に係るストレージシステム６００の外観構成を図５に示す。また、記憶デバイス制御装置１００の外観構成を図６に示す。

図５に示すように、本実施の形態に係るストレージシステム６００は記憶デバイス制御装置１００及び記憶デバイス３００がそれぞれの筐体に納められた形態をしている。記憶デバイス制御装置１００の筐体の両側に記憶デバイス３００の筐体が配置されている。

【００３６】

記憶デバイス制御装置１００は、正面中央部に管理端末１６０が備えられている。管理端末１６０はカバーで覆われており、図６に示すようにカバーを開けることにより管理端末１６０を使用することができる。なお図６に示した管理端末１６０はいわゆるノート型パーソナルコンピュータの形態をしているが、どのような形態とすることも可能である。

【００３７】

管理端末１６０の下部には、チャネル制御部１１０のボードを装着するためのスロットが設けられている。チャネル制御部１１０のボードとは、チャネル制御部１１０の回路基板が形成されたユニットであり、スロットへの装着単位である。本実施の形態に係るストレージシステム６００においては、スロットは８つあり、図５及び図６には８つのスロットにチャネル制御部１１０のボードが装着された状態が示されている。各スロットにはチャネル制御部１１０のボードを装着するためのガイドレールが設けられている。ガイドレールに沿ってチャネル制御部１１０のボードをスロットに挿入することにより、チャネル制御部１１０のボードを記憶デバイス制御装置１００に装着することができる。また各スロットに装着されたチャネル制御部１１０のボードは、ガイドレールに沿って手前方向に引き抜くことにより取り外すことができる。また各スロットの奥手方向正面部には、各チャネル制御部１１０のボードを記憶デバイス制御装置１００と電気的に接続するためのコネクタが設けられている。

【００３８】

なお、スロットには以上に説明したＮＡＳとして機能するタイプのチャネル制御部１１０以外にも、ＳＡＮ（Storage Area Network）に接続する機能を備えるタイプや、ＦＩＣＯＮ（Fibre Connection）（登録商標）やＥＳＣＯＮ(Enterprise System Connection)（登録商標）等のメインフレーム系のプロトコルに従って通信を行う機能を備えるタイプのチャネル制御部１１０が装着されることもある。またチャネル制御部１１０のボードを装着しないスロットを設けるようにすることもできる。

【００３９】

各スロットのチャネル制御部１１０は、同種の複数のチャネル制御部１１０でクラスタを構成する。例えば２枚のＣＨＮ１１０をペアとしてクラスタを構成することができる。クラスタを構成することにより、クラスタ内のあるチャネル制御部１１０に障害が発生した場合でも、障害が発生したチャネル制御部１１０がそれまで行っていた処理をクラスタ内の他のチャネル制御部１１０に引き継ぐようにすることができる（フェイルオーバ制御）。２枚のＣＨＮ１１０でクラスタ

を構成している様子を示す図を図１１に示すが、詳細は後述する。

【００４０】

なお、記憶デバイス制御装置１００は信頼性向上のため電源供給が２系統化されており、チャネル制御部１１０のボードが装着される上記８つのスロットは電源系統毎に４つずつに分けられている。そこでクラスタを構成する場合には、両方の電源系統のチャネル制御部１１０のボードを含むようにする。これにより、片方の電源系統に障害が発生し電力の供給が停止しても、同一クラスタを構成する他方の電源系統に属するチャネル制御部１１０のボードへの電源供給は継続されるため、当該チャネル制御部１１０に処理を引き継ぐ（フェイルオーバ）ことができる。

【００４１】

なお、上述したように、チャネル制御部１１０は上記各スロットに装着可能なボードとして提供されるが、上記一つのボードは一体形成された複数枚数の回路基板から構成されているようにすることもできる。

【００４２】

ディスク制御部１４０や共有メモリ１２０等の、記憶デバイス制御装置１００を構成する他の装置については図５及び図６には示されていないが、記憶デバイス制御装置１００の背面側等に装着されている。

また記憶デバイス制御装置１００には、チャネル制御部１１０のボード等から発生する熱を放出するためのファン１７０が設けられている。ファン１７０は記憶デバイス制御装置１００の上面部に設けられる他、チャネル制御部１１０用スロットの上部にも設けられている。

【００４３】

ところで、筐体に収容されて構成される記憶デバイス制御装置１００および記憶デバイス３００としては、例えばＳＡＮ対応として製品化されている従来構成の装置を利用することができる。特に上記のようにＣＨＮ１１０のボードのコネクタ形状を従来構成の筐体に設けられているスロットにそのまま装着できる形状とすることで従来構成の装置をより簡単に利用することができる。つまり本実施例のストレージシステム６００は、既存の製品を利用することで容易に構築する

ことができる。

【００４４】

＝＝＝チャネル制御部＝＝＝

本実施の形態に係るストレージシステム６００は、上述の通りＣＨＮ１１０により情報処理装置２００からのファイルアクセス要求を受け付け、ＮＡＳとしてのサービスを情報処理装置２００に提供する。

【００４５】

ＣＨＮ１１０のハードウエア構成を図７に示す。この図に示すようにＣＨＮ１１０のハードウエアは一体的にユニット化されたボードで構成される。以下、このユニットのことをＮＡＳボードとも記す。ＮＡＳボードは一枚もしくは複数枚の回路基板を含んで構成される。より具体的には、ＮＡＳボードは、ネットワークインタフェース部１１１、ＣＰＵ１１２、メモリ１１３、入出力制御部１１４、Ｉ／Ｏ（Input/Output）プロセッサ１１９、ＮＶＲＡＭ（Non Volatile RAM）１１５、ボード接続用コネクタ１１６、通信コネクタ１１７を備え、これらが同一のユニット化された回路基板として形成されて構成されている。

【００４６】

ネットワークインタフェース部１１１は、情報処理装置２００との間で通信を行うための通信インタフェースを備えている。ＣＨＮ１１０の場合は、例えばＴＣＰ／ＩＰプロトコルに従って情報処理装置２００から送信されたファイルアクセス要求を受信する。通信コネクタ１１７は情報処理装置２００と通信を行うためのコネクタである。ＣＨＮ１１０の場合はＬＡＮ４００に接続可能なコネクタであり、例えばイーサネット（登録商標）に対応している。

【００４７】

ＣＰＵ１１２は、ＣＨＮ１１０をＮＡＳボードとして機能させるための制御を司る。

メモリ１１３には様々なプログラムやデータが記憶される。例えば図８に示すメタデータ７３０やロックテーブル７２０、また図１０に示すＮＡＳマネージャ７０６等の各種プログラムが記憶される。

【００４８】

メタデータ７３０はファイルシステムプログラム７０３により実現されるファイルシステムが管理しているファイルに対応させて生成される情報である。メタデータ７３０には例えばファイルのデータが記憶されているＬＵ上のアドレスやデータサイズなど、ファイルの保管場所を特定するための情報が含まれる。メタデータ７３０にはファイルの容量、所有者、更新時刻等の情報が含まれることもある。また、メタデータ７３０はファイルだけでなくディレクトリに対応させて生成されることもある。メタデータ７３０の例を図１２に示す。メタデータ７３０は記憶デバイス３００上の各ＬＵにも記憶されている。

【００４９】

ロックテーブル７２０は、情報処理装置２００からのファイルアクセスに対して排他制御を行うためのテーブルである。排他制御を行うことにより情報処理装置２００でファイルを共用することができる。ロックテーブル７２０を図１３に示す。図１３に示すようにロックテーブル７２０にはファイルロックテーブル７２１とＬＵロックテーブル７２２とがある。ファイルロックテーブル７２１は、ファイル毎にロックが掛けられているか否かを示すためのテーブルである。いずれかの情報処理装置２００によりあるファイルがオープンされている場合に当該ファイルにロックが掛けられる。ロックが掛けられたファイルに対する他の情報処理装置２００によるアクセスは禁止される。ＬＵロックテーブル７２２は、ＬＵ毎にロックが掛けられているか否かを示すためのテーブルである。いずれかの情報処理装置２００により、あるＬＵに対するアクセスが行われている場合に当該ＬＵにロックが掛けられる。ロックが掛けられたＬＵに対する他の情報処理装置２００によるアクセスは禁止される。

【００５０】

入出力制御部１１４は、ディスク制御部１４０やキャッシュメモリ１３０、共有メモリ１２０、管理端末１６０との間でデータやコマンドの授受を行う。入出力制御部１１４はＩ／Ｏプロセッサ１１９やＮＶＲＡＭ１１５を備えている。Ｉ／Ｏプロセッサ１１９は例えば１チップのマイコンで構成される。Ｉ／Ｏプロセッサ１１９は上記データやコマンドの授受を制御し、ＣＰＵ１１２とディスク制御部１４０との間の通信を中継する。ＮＶＲＡＭ１１５はＩ／Ｏプロセッサ１１

９の制御を司るプログラムを格納する不揮発性メモリである。ＮＶＲＡＭ１１５に記憶されるプログラムの内容は、管理端末１６０や、後述するＮＡＳマネージャ７０６からの指示により書き込みや書き換えを行うことができる。

【００５１】

次にディスク制御部１４０のハードウエア構成を示す図を図９に示す。

ディスク制御部１４０は、一体的にユニット化されたボードとして形成されている。ディスク制御部１４０のボードは、インタフェース部１４１、メモリ１４３、ＣＰＵ１４２、ＮＶＲＡＭ１４４、ボード接続用コネクタ１４５を備え、これらが一体的にユニット化された回路基板として形成されている。

【００５２】

インタフェース部１４１は、接続部１５０を介してチャネル制御部１１０等との間で通信を行うための通信インタフェースや、記憶デバイス３００との間で通信を行うための通信インタフェースを備えている。

ＣＰＵ１４２は、ディスク制御部１４０全体の制御を司ると共に、チャネル制御部１４０や記憶デバイス３００、管理端末１６０との間の通信を行う。メモリ１４３やＮＶＲＡＭ１４４に格納された各種プログラムを実行することにより本実施の形態に係るディスク制御部１４０の機能が実現される。ディスク制御部１４０により実現される機能としては、記憶デバイス３００の制御やＲＡＩＤ制御、記憶デバイス３００に記憶されたデータの複製管理やバックアップ制御、リモートコピー制御等である。

【００５３】

ＮＶＲＡＭ１４４はＣＰＵ１４２の制御を司るプログラムを格納する不揮発性メモリである。ＮＶＲＡＭ１４４に記憶されるプログラムの内容は、管理端末１６０や、ＮＡＳマネージャ７０６からの指示により書き込みや書き換えを行うことができる。

またディスク制御部１４０のボードはボード接続用コネクタ１４５を備えている。ボード接続用コネクタ１４５が記憶デバイス制御装置１００側のコネクタと嵌合することにより、ディスク制御部１４０のボードは記憶デバイス制御装置１００と電気的に接続される。

【００５４】

＝＝＝ソフトウエア構成＝＝＝

次に、本実施の形態に係るストレージシステム６００におけるソフトウエア構成図を図１０に示す。

オペレーティングシステム７０１は例えばＵＮＩＸ（登録商標）である。オペレーティングシステム７０１上では、ＲＡＩＤマネージャ７０８、ボリュームマネージャ７０７、ＳＶＰマネージャ７０９、ファイルシステムプログラム７０３、ネットワーク制御部７０２、障害管理プログラム７０５、ＮＡＳマネージャ７０６などのソフトウエアが動作する。

【００５５】

オペレーティングシステム７０１上で動作するＲＡＩＤマネージャ７０８は、ＲＡＩＤ制御部７４０に対するパラメータの設定やＲＡＩＤ制御部７４０を制御する機能を提供する。ＲＡＩＤマネージャ７０８はオペレーティングシステム７０１やオペレーティングシステム７０１上で動作する他のアプリケーション、もしくは管理端末１６０からパラメータや制御指示情報を受け付けて、受け付けたパラメータのＲＡＩＤ制御部７４０への設定や、ＲＡＩＤ制御部指示情報に対応する制御コマンドの送信を行う。

【００５６】

ここで設定されるパラメータとしては、例えば、ＲＡＩＤグループを構成する記憶デバイス（物理ディスク）を定義（ＲＡＩＤグループの構成情報、ストライプサイズの指定など）するためのパラメータ、ＲＡＩＤレベル（例えば０，１，５）を設定するためのパラメータなどがある。また、ＲＡＩＤマネージャ７０８がＲＡＩＤ制御部７４０に送信する制御コマンドとしてはＲＡＩＤの構成・削除・容量変更を指示するコマンド、各ＲＡＩＤグループの構成情報を要求するコマンドなどがある。

【００５７】

ボリュームマネージャ７０７は、ＲＡＩＤ制御部７４０によって提供されるＬＵをさらに仮想化した仮想化論理ボリュームをファイルシステムプログラム７０３に提供する。１つの仮想化論理ボリュームは１以上の論理ボリュームによって

構成される。

【００５８】

ファイルシステムプログラム７０３の主な機能は、ネットワーク制御部７０２が受信したファイルアクセス要求に指定されているファイル名とそのファイル名が格納されている仮想化論理ボリューム上のアドレスとの対応づけを管理することである。例えば、ファイルシステムプログラム７０３はファイルアクセス要求に指定されているファイル名に対応する仮想化論理ボリューム上のアドレスを特定する。

【００５９】

ネットワーク制御部７０２は、ＮＦＳ（Network File System）７１１とＣＩＦＳ（Common Interface File System）７１３の２つのファイルシステムプロトコルを含んで構成される。ＮＦＳ７１１は、ＮＦＳ７１１が動作するＵＮＩＸ（登録商標）系の情報処理装置２００からのファイルアクセス要求を受け付ける。一方、ＣＩＦＳ７１３はＣＩＦＳ７１３が動作するＷｉｎｄｏｗｓ（登録商標）系の情報処理装置２００からのファイルアクセス要求を受け付ける。

【００６０】

ＮＡＳマネージャ７０６は、ストレージシステム６００について、その動作状態の確認、設定や制御などを行うためのプログラムである。ＮＡＳマネージャ７０６はＷｅｂサーバとしての機能も有し、情報処理装置２００からストレージシステム６００の設定や制御を行うための設定Ｗｅｂページを情報処理装置２００に提供する。設定Ｗｅｂページはチャネル制御部１１０の個々において動作するＮＡＳマネージャ７０６により提供される。ＮＡＳマネージャ７０６は、情報処理装置２００からのＨＴＴＰ（HyperText Transport Protocol）リクエストに応じて、設定Ｗｅｂページのデータを情報処理装置２００に送信する。情報処理装置２００に表示された設定Ｗｅｂページを利用してシステムアドミニストレータなどによりストレージシステム６００の設定や制御の指示が行われる。

【００６１】

ＮＡＳマネージャ７０６は、設定Ｗｅｂページに対する操作に起因して情報処理装置２００から送信される設定や制御に関するデータを受信してそのデータに

対応する設定や制御を実行する。これにより、情報処理装置２００からストレージシステム６００の様々な設定や制御を行うことができる。またＮＡＳマネージャ７０６は設定Ｗｅｂページの設定内容をチャネル制御部１１０上で動作するＯＳやアプリケーションプログラム、ディスク制御部１４０等に通知する。設定Ｗｅｂページで設定された内容は共有ＬＵ３１０に管理されることもある。

【００６２】

ＮＡＳマネージャ７０６の設定Ｗｅｂページを利用して行うことができる内容としては、例えば、ＬＵの管理や設定（容量管理や容量拡張・縮小、ユーザ割り当て等）、上述の複製管理やリモートコピー（レプリケーション）等の機能に関する設定や制御（複製元のＬＵと複製先のＬＵの設定など）、冗長構成されたＣＨＮでのクラスタの管理（フェイルオーバさせる相手の対応関係の設定、フェイルオーバ方法など）、ＯＳやＯＳ上で動作するアプリケーションプログラムのバージョン管理などがある。

【００６３】

なお、ＮＡＳマネージャ７０６によるストレージシステム６００の動作状態の確認、設定や制御には、上述した設定Ｗｅｂページを介する方法以外にクライアント・サーバシステムとすることもできる。その場合、ＮＡＳマネージャ７０６はクライアント・サーバシステムのサーバ機能を有し、情報処理装置２００のクライアント機能の操作に起因して送信される設定や制御を、上述した設定Ｗｅｂページと同様に実施することで、ストレージシステム６００の動作状態の確認、設定や制御を行う。

【００６４】

ＳＶＰマネージャ７０９は、管理端末１６０からの要求に応じて各種のサービスを管理端末１６０に提供する。例えば、ＬＵの設定内容やＲＡＩＤの設定内容等のストレージシステム６００に関する各種設定内容の管理端末１６０への提供や、管理端末１６０から入力されたストレージシステム６００に関する各種設定の反映等を行う。

【００６５】

＝＝＝クラスタ間同期制御＝＝＝

障害管理プログラム７０５は、クラスタを構成するチャネル制御部１１０間でのフェイルオーバ制御を行うためのプログラムである。

２枚のＣＨＮ１１０でクラスタ１８０が構成されている様子を示す図を図１１に示す。図１１では、ＣＨＮ１（チャネル制御部１）１１０とＣＨＮ２（チャネル制御部２）１１０とでクラスタ（グループ）１８０が構成されている場合を示す。

上述したように、フェイルオーバ処理はクラスタ１８０を構成するチャネル制御部１１０間で行われる。例えばＣＨＮ１（１１０）に何らかの障害が発生し、処理を継続することができなくなった場合には、ＣＨＮ１（１１０）がそれまで行っていた処理はＣＨＮ２（１１０）に引き継がれ、ＣＨＮ２（１１０）がその後の処理を行う。

【００６６】

なお、フェイルオーバはＣＨＮ１１０に障害が発生したときに自動的に行われることもあるが、オペレータが管理端末１６０を操作して手動で行われることもある。またユーザがＮＡＳマネージャ７０６により提供される設定Ｗｅｂページを利用することにより、情報処理装置２００から手動で行われることもある。フェイルオーバを手動で行う目的としては、耐用年数の経過やバージョンアップ、定期診断などのためにチャネル制御部１１０のハードウエア（例えばＮＡＳボード）を交換する必要が生じた場合などがある。

【００６７】

ＣＨＮ２（１１０）がＣＨＮ１（１１０）の処理を引き継いで実行するためには、ＣＨＮ２（１１０）はＣＨＮ１（１１０）から種々のデータを引き継ぐことが必要である。ＣＨＮ１（１１０）からＣＨＮ２（１１０）に引き継がれるデータは、例えばＮＦＳユーザデータ、ＣＩＦＳユーザデータ、装置管理者データ、フェイルオーバ用ハートビート、ＣＨＮ１（１１０）のＩＰアドレス、ＮＦＳファイルロック情報、クラスタ制御情報等である。

【００６８】

ＮＦＳユーザデータは、ＵＮＩＸ（登録商標）系オペレーティングシステムが実行される情報処理装置２００を利用してＣＨＮ１（１１０）によりファイルア

クセスサービスの提供を受けることができるユーザを管理するためのデータである。例えばユーザのログインＩＤやパスワード等が登録されるデータである。ＣＨＮ２（１１０）はＣＨＮ１（１１０）のＮＦＳユーザデータを引き継ぐことにより、それまでＣＨＮ１（１１０）によりファイルアクセスサービスの提供を受けていたユーザに対して、継続してファイルアクセスサービスを提供することが可能となる。

【００６９】

ＣＩＦＳユーザデータは、Ｗｉｎｄｏｗｓ（登録商標）系オペレーティングシステムが実行される情報処理装置２００を利用してＣＨＮ１（１１０）によりファイルアクセスサービスの提供を受けることができるユーザを管理するためのデータである。例えばユーザのログインＩＤやパスワード等が登録されるデータである。ＣＨＮ２（１１０）はＣＨＮ１（１１０）のＣＩＦＳユーザデータを引き継ぐことにより、それまでＣＨＮ１（１１０）によりファイルアクセスサービスの提供を受けていたユーザに対して、継続してファイルアクセスサービスを提供することが可能となる。

【００７０】

装置管理者データは、ストレージシステム６００あるいは記憶デバイス制御装置１００の管理者を管理するためのデータである。例えば管理者のログインＩＤやパスワード、ホームディレクトリの位置を示すデータを含む。装置管理者データは、クラスタ１８０に関係なく、記憶デバイス制御装置１００内の全チャネル制御部１１０に共通のデータである。

【００７１】

フェイルオーバ用ハートビートは、クラスタ１８０内の各ＣＨＮ１１０が相互に動作状態を確認し合うためのデータである。ＣＨＮ１（１１０）及びＣＨＮ２（１１０）は、相互に自己の処理が正常に行われていることを示すデータ（ハートビートマーク）を定期的に通知する。そして、相手側の通知有無を確認する。相手側による通知が確認できない場合には、相手側に何らかの障害が発生したと判断する。ハートビートマークには、ＣＨＮ１１０の識別子や、ＣＨＮ１１０が正常であるか異常であるかを示す符号、更新時刻等の情報が含まれる。

【００７２】

ＣＨＮ１１０のＩＰアドレスは、ＬＡＮ４００上でＴＣＰ／ＩＰ通信プロトコルに従って通信を行う場合にＣＨＮ１１０を識別するためのアドレスである。例えばＣＨＮ１（１１０）のＩＰアドレスをＣＨＮ２（１１０）が引き継ぐことにより、それまでＣＨＮ１（１１０）がＬＡＮ４００を通じて受信していたデータをＣＨＮ２（１１０）が受信することができるようになる。

【００７３】

ＮＦＳファイルロック情報は、ファイルロックテーブル７２１やＬＵロックテーブル７２２を含むファイルのロック状態を管理するためのデータである。

クラスタ制御情報は、クラスタ内のＣＨＮ１１０間で引き継ぎが必要な上記以外のデータを含む。例えば、障害が発生したＣＨＮ１１０が管理していたＬＵに構築されているファイルシステムのマウントに関するマウント情報やネットワークインタフェース部１１１のＭＡＣ（Media Access Control）アドレスやネットワークファイルシステムのエクスポート情報である。

ＣＨＮ２（１１０）は、これらの引き継ぎデータをＣＨＮ１（１１０）から引き継ぐことにより、ＣＨＮ１（１１０）がそれまで行っていた処理を引き継いで行う。

【００７４】

本実施の形態における記憶デバイス制御装置１００では、引き継ぎ元のＣＨＮ１１０と引き継ぎ先のＣＨＮ１１０との間で、これらの引き継ぎデータの同期を取ることにより引き継ぎが行われる。すなわち、引き継ぎ元のＣＨＮ１１０と引き継ぎ先のＣＨＮ１１０との間では、引き継ぎデータの内容が同一となるように制御される。

例えば、あるＣＨＮ１１０により上記いずれかの引き継ぎデータが更新された場合には、更新された引き継ぎデータを、ＣＨＮ１１０間を相互に接続するネットワークを介して他のＣＨＮ１１０に送信する。このようにすることにより、引き継ぎ元のＣＨＮ１１０と引き継ぎ先のＣＨＮ１１０とで参照される引き継ぎデータの内容を同一にすることができる。ここで、ネットワークとしてはＬＡＮ４００を用いるようにすることもできるし、接続部１５０を用いるようにすること

もできる。また内部ＬＡＮ１５１を用いるようにすることもできる。

【００７５】

また複数のＣＨＮ１１０から共通にアクセス可能な共有ＬＵ（共有ボリューム）に引き継ぎデータを記憶するようにすることにより引き継ぎデータの同期を実現することもできる。これにより、引き継ぎ元のＣＨＮ１１０と引き継ぎ先のＣＨＮ１１０とで同一の引き継ぎデータが参照されるようにできる。

また、複数のＣＨＮ１１０から共通にアクセス可能な共有メモリ１２０に引き継ぎデータを記憶するようにすることにより、引き継ぎ元のＣＨＮ１１０と引き継ぎ先のＣＨＮ１１０とで同一の引き継ぎデータが参照されるようにすることもできる。

【００７６】

本実施の形態における記憶デバイス制御装置１００で行われる、引き継ぎデータの同期を説明するためのシステム構成図を図１４及び図１５に示す。

図１４及び図１５に示す記憶デバイス制御装置１００では、ＣＨＮ１（１１０）とＣＨＮ２（１１０）とでクラスタＡ（１８０）が構成され、ＣＨＮ３（１１０）とＣＨＮ４（１１０）とでクラスタＢ（１８０）が構成されている。つまり、ＣＨＮ１乃至４（１１０）はクラスタＡ（１８０）とクラスタＢ（１８０）とに類別されている。そして各ＣＨＮ１１０はＬＡＮ４００を介して情報処理装置２００と接続されると共に、ＣＨＮ１１０間も相互に接続される。また各ＣＨＮ１１０は接続部１５０を介して共有メモリ１２０、システムＬＵ、ユーザＬＵ、管理端末１６０と接続されている。図１４においては、システムＬＵはＣＨＮ１１０毎に割り当てられていることが示されている。従って図１４に示すシステムＬＵはローカルＬＵでもある。なお、図１４及び図１５に示すＮＡＳ制御ソフトウエアは、障害管理プログラム７０５を含む。

【００７７】

図１５は、ストレージシステム６００が備える各ＬＵを、ローカルＬＵ（ローカルボリューム）、共有ＬＵ（共有ボリューム）、全体共有ＬＵ（第２の共有ボリューム）に分類して記載したものである。ローカルＬＵは各ＣＨＮ１１０が個別にアクセス可能なＬＵである。共有ＬＵはクラスタ内の複数のＣＨＮ１１０か

ら共通にアクセス可能なＬＵである。全体共有ＬＵは、ストレージシステム６００内の全ＣＨＮ１１０で共通にアクセス可能なＬＵである。

【００７８】

上述のように、フェイルオーバのための引き継ぎデータにはＮＦＳユーザデータのようにＣＨＮ１１０毎に個別に作成されるデータもあるし、装置管理者データのようにストレージシステム６００全体で共通なデータもある。そのため本実施の形態に係る記憶デバイス制御装置１００においては、引き継ぎデータの種類に応じて、異なる方法により同期が取られる。本実施の形態におけるフェイルオーバ時に引き継がれるデータと同期の方法を示す同期方法管理テーブルを図１６に示す。同期方法管理テーブルは、各ＣＨＮ１１０のメモリ１１３に記憶しておくようにすることもできるし、共有メモリ１２０に記憶しておくようにすることもできる。また、各ＣＨＮ１１０のローカルＬＵに記憶しておくようにすることもできる。

【００７９】

図１６に示す同期方法管理テーブルは、”制御情報欄”、”データの同期種別欄”、”同期方法欄”、”同期データ欄”、”ローカルＬＵへの反映要否欄”を含んで構成される。

”制御情報欄”には、引き継ぎデータの種類が記載される。上述したように本実施の形態においては、ＮＦＳユーザデータ、ＣＩＦＳユーザデータ、装置管理者データ、フェイルオーバ用ハートビート、ＣＨＮ（１１０）のＩＰアドレス、ＮＦＳファイルロック情報、クラスタ制御情報が記載される。なお以下、引き継ぎデータのことを制御情報とも記す。

【００８０】

”データの同期種別欄”には、各引き継ぎデータの同期を取る範囲が記載される。「クラスタ内で同期」と記載されている場合は、クラスタ内で当該引き継ぎデータの同期が取られることを示す。すなわち、更新された引き継ぎデータはクラスタ内の他のＣＨＮ１１０との間で同期が取られる。「ストレージ装置で同期」と記載されている場合には、ストレージシステム６００全体で当該引き継ぎデータの同期が取られることを示す。「システム固有」と記載されている場合は、当

該引き継ぎデータは更新されることがないため同期を取る必要がないことを示す

”同期方法欄”には、各引き継ぎデータの同期をとる方法が記載される。「ネットワーク」と記載されている場合は、あるＣＨＮ１１０により更新された引き継ぎデータは、ＣＨＮ１１０間を相互に接続するネットワークを介して他のＣＨＮ１１０に送信される。ネットワークとしてはＬＡＮ４００を用いるようにすることもできるし、接続部１５０を用いるようにすることもできる。また内部ＬＡＮ１５１を用いるようにすることもできる。「共有ＬＵ」と記載されている場合は、あるＣＨＮ１１０により更新された引き継ぎデータは、共有ＬＵに記憶される。「共有メモリ」と記載されている場合は、あるＣＨＮ１１０により更新された引き継ぎデータは、共有メモリに記憶される。「－」と記載されている場合は、同期不要であることを示す。本実施の形態においては、装置管理者データ及びＣＨＮのＩＰアドレスは更新されることがないので同期不要としているが、同期をとるようにすることも可能である。

【００８１】

”同期データ欄”には、各引き継ぎデータに関するコメント等が記載される。例えば引き継ぎデータの具体的なファイル名等を記載しておくことができる。同期方法管理テーブルにおいて”同期データ欄”を設けない構成とすることもできる。

【００８２】

”ローカルＬＵへの反映要否欄”は、更新された引き継ぎデータを共有ＬＵや共有メモリ１２０に記載することにより同期を取った場合に、当該共有ＬＵや共有メモリ１２０を共通にアクセス可能な他のＣＨＮ１１０のローカルＬＵにも当該引き継ぎデータを書き込むようにするか否かを選択するための欄である。「否」と記載されている場合は、更新された引き継ぎデータは他のＣＨＮ１１０のローカルＬＵには書き込まれない。従ってこの場合他のＣＨＮ１１０は、共有ＬＵあるいは共有メモリ１２０にアクセスすることにより当該引き継ぎデータを参照する。「要」と記載されている場合は、更新された引き継ぎデータは他のＣＨＮ１１０のローカルＬＵにも書き込まれる。従ってこの場合、他のＣＨＮ１１０は自己のローカルＬＵにアクセスすることにより当該引き継ぎデータを参照することができる。

【００８３】

例えば、更新頻度は少ないが参照頻度が多い引き継ぎデータは、共有ＬＵや共有メモリ１２０の他、ローカルＬＵにも記憶するようにしておく。これにより、共有ＬＵや共有メモリ１２０へのアクセス頻度を減少させ、他のＣＨＮ１１０との間で発生するアクセス競合を減少させることができるので、データアクセス性能を向上させることができる。また、一時的にしか参照されない引き継ぎデータや頻繁に更新される引き継ぎデータなどはローカルＬＵに反映させないようにする。これにより引き継ぎデータをローカルＬＵへ反映させる際の処理オーバヘッドを小さくすることができる。

このように、本実施の形態に係る記憶デバイス制御装置１００においては、フェイルオーバの引き継ぎデータの種類に応じた特性を考慮して、最適な方法により同期を取ることができる。

【００８４】

次に、本実施の形態に係るフェイルオーバ時に引き継がれるデータの同期の方法を決定するための処理を示すフローチャートを図１７に示す。なお以下の処理は、本実施の形態に係る各種の動作を行うためのコードから構成される障害管理プログラム７０５を、ＣＰＵ１１２が実行することにより実現される。

【００８５】

まず制御情報が発生する（S1000）。ここで制御情報が発生するとはＣＨＮ１１０内の他のプログラムや管理端末１６０内のプログラム、あるいは情報処理装置２００内のプログラムから、フェイルオーバのための引き継ぎデータの更新要求を受領したという意味である。例えばファイルアクセスサービスの提供を受けるＮＦＳユーザの追加や削除を行うために更新されたＮＦＳユーザデータを情報処理装置２００から受信した場合や、クラスタ１８０内のＣＨＮ１１０により定期的に更新されるハートビートマークの更新要求を受領した場合である。

【００８６】

これらの引き継ぎデータの更新は自動的に行われることもあるが、オペレータが管理端末１６０を操作して手動で行われることもある。またユーザがＮＡＳマネージャ７０６により提供される設定Ｗｅｂページを利用することにより、情報

処理装置２００から手動で行われることもある。自動的に行われる場合としては例えばハートビートマークが更新される場合があり、手動で行われる場合としてはＮＦＳユーザデータが更新される場合がある。

【００８７】

次に、ＣＰＵ１１２は同期方法管理テーブルの”データの同期種別欄”を参照し、当該引き継ぎデータが他のＣＨＮ１１０やストレージシステム６００全体で利用されるデータであるかを確認する（S1001）。当該引き継ぎデータが他のＣＨＮ１１０と同期を必要としないデータである場合には、自己のローカルＬＵに当該引き継ぎデータを書き込んで終了する（S1002）。

【００８８】

一方、当該引き継ぎデータが他のＣＨＮ１１０と同期を必要とするデータである場合には、同期方法管理テーブルの”データの同期種別欄”を参照することにより、クラスタ内での同期が必要なデータであるかどうかを確認する（S1003）。

【００８９】

クラスタ内での同期が必要ではないデータである場合には、当該データはストレージシステム６００全体での同期が必要なデータであるので、接続部１５０を介して当該データを全体共有ＬＵに書き込む（S1004）。これにより、当該引き継ぎデータはストレージシステム６００内の全てのＣＨＮ１１０から同一の内容で参照されるようにすることができる。そして同期方法管理テーブルの”ローカルＬＵへの反映要否欄”を参照することにより、当該引き継ぎデータをローカルＬＵにも反映するかどうかを確認する（S1005）。同期方法管理テーブルの”ローカルＬＵへの反映要否欄”が「否」の場合はそのまま処理を終了する。”ローカルＬＵへの反映要否欄”が「要」の場合は、S1004において全体共有ＬＵに書き込んだ引き継ぎデータを、他のＣＨＮ１１０のローカルＬＵにも書き込む。この場合にはストレージシステム６００内の全ＣＨＮ１１０の各ローカルＬＵに引き継ぎデータを書き込む（S1006）。これにより各ＣＨＮ１１０は、引き継ぎデータを参照する際には自己のローカルＬＵにアクセスすればよく、全体共有ＬＵにアクセスする必要がなくなる。全体共有ＬＵにアクセスする必要がなくなることにより他のＣＨＮ１１０とのアクセス競合の頻度を減少させることができるので、デ

ータアクセス性能を向上させることができる。

【００９０】

一方、S1003においてクラスタ内での同期が必要なデータであることが判明した場合には、クラスタ構成のＣＨＮ１１０を調査して、通知先ＣＨＮ１１０を特定する（S1007）。すなわち自己と同一のクラスタに属する他のＣＨＮ１１０を特定する。特定は例えばＣＨＮ１１０の識別子により行うことができる。続いて、引き継ぎデータの同期方法にネットワークを利用するかどうかを同期方法管理テーブルの”同期方法欄”を参照することにより確認する（S1008）。”同期方法欄”の記載が「ネットワーク」以外の場合は、共有ＬＵを使用して同期するデータであるかどうかを確認する（S1011）。同期方法管理テーブルの”同期方法欄”の記載が「共有ＬＵ」の場合には”Ｙ”に進み、クラスタ内の共有ＬＵに引き継ぎデータを書き込む（S1012）。これにより、当該引き継ぎデータは自己と同一クラスタ１８０内の全てのＣＨＮ１１０から同一の内容で参照されるようにすることができる。

【００９１】

そして必要に応じて、クラスタ内の他のＣＨＮ１１０に対して上記引き継ぎデータの共有ＬＵ内での記憶位置を通知するようにする（S1013）。すなわち、当該引き継ぎデータのメタデータをクラスタ内の他のＣＨＮ１１０に対して通知する。これにより、クラスタ内の他のＣＨＮ１１０は当該引き継ぎデータを参照する際に、共有ＬＵ内のどこに引き継ぎデータが格納されているのかを知ることができる。なお引き継ぎデータによっては記憶位置が特定の場所に固定されているものもある。このようなデータの場合には、引き継ぎデータの共有ＬＵにおける記憶位置を他のＣＨＮ１１０に通知する必要はない。

【００９２】

次に同期方法管理テーブルの”ローカルＬＵへの反映要否欄”を参照することにより、当該引き継ぎデータをローカルＬＵにも反映するかどうかを確認する（S1014）。同期方法管理テーブルの”ローカルＬＵへの反映要否欄”が「否」の場合はそのまま処理を終了する。”ローカルＬＵへの反映要否欄”が「要」の場合は、S1012において共有ＬＵに書き込んだ引き継ぎデータを、クラスタ内の他のＣＨ

Ｎ１１０のローカルＬＵにも書き込む（S1015）。これによりクラスタ内の各ＣＨＮ１１０は、引き継ぎデータを参照する際には自己のローカルＬＵにアクセスすればよく、共有ＬＵにアクセスする必要がなくなる。共有ＬＵにアクセスする必要がなくなることにより他のＣＨＮ１１０とのアクセス競合の頻度を減少させることができるので、データアクセス性能を向上させることができる。

【００９３】

一方S1011において、同期方法管理テーブルの”同期方法欄”の記載が「共有メモリ」の場合には”N”に進み、共有メモリ１２０に引き継ぎデータを書き込む（S1016）。これにより、当該引き継ぎデータは自己と同一クラスタ１８０内の全てのＣＨＮ１１０から同一の内容で参照されるようにすることができる。

【００９４】

そして必要に応じて、クラスタ内の他のＣＨＮ１１０に対して上記引き継ぎデータの共有メモリ１２０内での記憶位置を通知するようにする（S1017）。すなわち、当該引き継ぎデータのメタデータをクラスタ内の他のＣＨＮ１１０に対して通知する。これにより、クラスタ内の他のＣＨＮ１１０は当該引き継ぎデータを参照する際に、共有メモリ１２０内のどこに引き継ぎデータが格納されているのかを知ることができる。また、引き継ぎデータによっては記憶位置が特定のアドレスに固定されているものもある。このようなデータの場合には、引き継ぎデータの共有メモリ１２０における記憶位置を他のＣＨＮ１１０に通知する必要はない。

【００９５】

次に同期方法管理テーブルの”ローカルＬＵへの反映要否欄”を参照することにより、当該引き継ぎデータをローカルＬＵにも反映するかどうかを確認する（S1018）。同期方法管理テーブルの”ローカルＬＵへの反映要否欄”が「否」の場合はそのまま処理を終了する。”ローカルＬＵへの反映要否欄”が「要」の場合は、S1016において共有メモリ１２０に書き込んだ引き継ぎデータを、クラスタ内の他のＣＨＮ１１０のローカルＬＵにも書き込む（S1015）。これによりクラスタ内の各ＣＨＮ１１０は、引き継ぎデータを参照する際には自己のローカルＬＵにアクセスすればよく、共有メモリ１２０にアクセスする必要がなくなる。共有メ

モリ１２０にアクセスする必要がなくなることにより他のＣＨＮ１１０とのアクセス競合の頻度を減少させることができるので、データアクセス性能を向上させることができる。

【００９６】

一方S1008において、同期方法管理テーブルの”同期方法欄”の記載が「ネットワーク」の場合は”Ｙ”にすすむ。まず引き継ぎデータを自己のローカルＬＵに書き込む（S1009）。そして、自己のローカルＬＵに書き込んだのと同じ引き継ぎデータをクラスタ内の他のＣＨＮ１１０に対してネットワークを介して送信する（S1010）。これにより当該引き継ぎデータがクラスタ内の他のＣＨＮ１１０のローカルＬＵにも反映され、自己のローカルＬＵに記憶されている引き継ぎデータの内容と、クラスタ内の他のＣＨＮ１１０のローカルＬＵに記憶されている引き継ぎデータの内容とを同一にすることができる。

【００９７】

このように、本実施の形態に係る記憶デバイス制御装置１００においては、フェイルオーバの引き継ぎデータの種類に応じた特性を考慮して、最適な方法により同期処理を行うことができる。また引き継ぎデータの同期を取るようにすることにより、ＣＨＮ１１０に障害が発生してからデータの引き継ぎを行う必要がなくなり、フェイルオーバを迅速に行うことが可能となる。

【００９８】

次に上記引き継ぎデータを参照する際の処理について説明する。引き継ぎデータの参照の処理は、本実施の形態に係る各種の動作を行うためのコードから構成される障害管理プログラム７０５を、ＣＰＵ１１２が実行することにより実現される。ＣＰＵ１１２は、図１８に示す引き継ぎデータ参照テーブルを参照することにより引き継ぎデータの格納先（参照先）を知ることができる。引き継ぎデータ参照テーブルは、各ＣＨＮ１１０のメモリ１１３に記憶しておくようにすることもできるし、共有メモリ１２０に記憶しておくようにすることもできる。また、各ＣＨＮ１１０のローカルＬＵに記憶しておくようにすることもできる。

【００９９】

図１８に示す引き継ぎデータ参照テーブルは、”制御情報欄”、”データの格納

先欄”、”データの通知有無欄”を含んで構成される。

”制御情報欄”には、引き継ぎデータの種類が記載される。本実施の形態においては、ＮＦＳユーザデータ、ＣＩＦＳユーザデータ、装置管理者データ、フェイルオーバ用ハートビート、ＮＦＳファイルロック情報、クラスタ制御情報が記載される。

【０１００】

”データの格納先欄”には、当該引き継ぎデータの格納先（参照先）が記載されている。「ローカルＬＵ」と記載されている場合には、当該引き継ぎデータは自己のローカルＬＵに記憶されていることを示す。すなわち、他のＣＨＮ１１０が当該引き継ぎデータを更新した際に、当該引き継ぎデータがネットワークを介して送信されてきたか、又は、他のＣＨＮ１１０により自己のローカルＬＵにも当該引き継ぎデータが書き込まれたかのいずれかにより、当該引き継ぎデータは自己のローカルＬＵに記憶されていることを示す。「共有ＬＵ」と記載されている場合には、当該引き継ぎデータは自己が属するクラスタ１８０内のＣＨＮ１１０で共有している共有ＬＵに記憶されていることを示す。「全体共有ＬＵ」と記載されている場合には、当該引き継ぎデータはストレージシステム６００内の全ＣＨＮ１１０で共有している全体共有ＬＵに記憶されていることを示す。「共有メモリ」と記載されている場合には、当該引き継ぎデータは共有メモリ１２０に記憶されていることを示す。

【０１０１】

”データの通知有無欄”には、当該引き継ぎデータを更新した他のＣＨＮ１１０から、引き継ぎデータの記憶位置の通知を受けたか否かが記載される。「あり」と記載されている場合には通知を受けたことが示され、「なし」と記載されている場合には通知を受けていないことが示される。「－」と記載されている場合は通知の有無は関係ないことが示される。ＮＦＳユーザデータの場合は図１６の同期方法管理テーブルにより、ネットワーク経由で他のＣＨＮ１１０から送信されてくる。そのため、当該ＮＦＳユーザデータをローカルＬＵに記憶するのは自ＣＨＮ１１０であり、他のＣＨＮ１１０から記憶位置の通知は行われないからである。

【０１０２】

このように、本実施の形態に係る記憶デバイス制御装置１００においては、引き継ぎデータ参照テーブルを参照することにより引き継ぎデータの格納先を知ることができる。

【０１０３】

次に本実施の形態に係るフェイルオーバ時に引き継がれるデータを参照する際の処理を示すフローチャートを図１９に示す。

まず引き継ぎデータの参照要求を受領する（S2000）。引き継ぎデータの参照要求は、ＣＨＮ１１０内の他のプログラムや管理端末１６０内のプログラム、あるいは情報処理装置２００内のプログラムから受領する。例えばファイルアクセスサービスの提供を受けるＮＦＳユーザの追加や削除を行うために情報処理装置２００から要求を受ける場合や、クラスタ１８０内の他のＣＨＮ１１０のハートビートマークを確認するために参照要求を受ける場合がある。

【０１０４】

次に引き継ぎデータ参照テーブルの”データの格納先欄”を参照することにより、当該引き継ぎデータがローカルＬＵに記憶されているか否かを確認する（S2001）。”データの格納先欄”に「ローカルＬＵ」と記載されている場合には、自己のローカルＬＵにアクセスして当該引き継ぎデータを参照する（S2002）。引き継ぎデータの記憶位置はメタデータを参照することにより知ることができる。

【０１０５】

”データの格納先欄”に「ローカルＬＵ」以外が記載されている場合には、引き継ぎデータは共有ＬＵ、共有メモリ、又は全体共有ＬＵのいずれかに記載されている。そこでまず、引き継ぎデータ参照テーブルの”データの通知有無欄”を参照することにより、他のＣＨＮ１１０から当該引き継ぎデータに関する通知を受けているか否かを確認する（S2003）。

【０１０６】

通知を受けていない場合には、当該引き継ぎデータは共有ＬＵ、共有メモリ、あるいは全体共有ＬＵのいずれかの所定の記憶位置に記憶されている。そのため、ＣＰＵ１１２は、定期的にこれらの所定の記憶位置を参照し、引き継ぎデータ

の更新がなされていないかを確認するようにする。なお上記の所定の記憶位置は、引き継ぎデータ参照テーブルに記録されるようにしておくこともできるし、引き継ぎデータ参照テーブルとは別のテーブルに記憶しておくようにすることもできる。

【０１０７】

S2004において一定時間の経過を待った後、引き継ぎデータ参照テーブルの”データの格納先欄”を参照することにより、クラスタ内の共有ＬＵに当該引き継ぎデータが記憶されているかどうかを確認する（S2007）。”データの格納先欄”に「共有ＬＵ」と記憶されている場合には、共有ＬＵの所定の記憶位置にアクセスして当該引き継ぎデータを読み出す（S2008）。次に、引き継ぎデータ参照テーブルの”データの格納先欄”を参照することにより、全体共有ＬＵに当該引き継ぎデータが記憶されているかどうかを確認する（S2009）。”データの格納先欄”に「全体共有ＬＵ」と記憶されている場合には、全体共有ＬＵの所定の記憶位置にアクセスして当該引き継ぎデータを読み出す（S2010）。”データの格納先欄”に「共有メモリ」と記憶されている場合には、共有メモリの所定の記憶位置にアクセスして当該引き継ぎデータを読み出す（S2011）。

【０１０８】

一方、S2003において他のＣＨＮ１１０から当該引き継ぎデータに関する通知を受けている場合には、その通知には当該引き継ぎデータの記憶位置が指定してあるか否かを確認する（S2005）。もし、当該引き継ぎデータの記憶位置が指定してある場合には、共有メモリ、共有ＬＵ、あるいは全体共有ＬＵの指定された記憶位置から当該引き継ぎデータを読み出す（S2006）。

【０１０９】

もし、当該引き継ぎデータの記憶位置が指定されてない場合には、当該引き継ぎデータは共有ＬＵ、共有メモリ、あるいは全体共有ＬＵのいずれかの所定の記憶位置に記憶されている。そこで引き継ぎデータ参照テーブルの”データの格納先欄”を参照することにより、クラスタ内の共有ＬＵに当該引き継ぎデータが記憶されているかどうかを確認する（S2007）。以下、上述の処理と同様の処理を行う。

【０１１０】

このように本実施の形態における記憶デバイス制御装置１００においては、引き継ぎデータ参照テーブルを参照しながら上記の処理が行われることにより、その種類に応じてさまざまな位置に記憶されているフェイルオーバの引き継ぎデータを、適切に読み出すことができる。

【０１１１】

次に、本実施の形態に係るフェイルオーバ制御を示すフローチャートを図２０に示す。フェイルオーバ制御は、各種の動作を行うためのコードから構成される障害管理プログラム７０５をＣＨＮ１１０が備えるＣＰＵ１１２が実行することにより実現される。

【０１１２】

図１１で示したようにフェイルオーバ制御はクラスタ１８０を構成するＣＨＮ１１０間で行われる。フェイルオーバ制御はＣＨＮ１１０に障害が発生した場合の他、ＮＡＳマネージャ７０６からの指示（フェイルオーバの実行要求）によっても行われるが、図２０においては、ＣＨＮ１（１１０）とＣＨＮ２（１１０）とで構成されるクラスタ内において、ＣＨＮ１（１１０）で異常が発生した場合のフェイルオーバ制御の例を示す。

【０１１３】

まずＮＦＳ／ＣＩＦＳファイル共有データがユーザによって追加される（S3000）。ＮＦＳ／ＣＩＦＳファイル共有データとは、ＵＮＩＸ（登録商標）系オペレーティングシステムが実行される情報処理装置２００、又はＷｉｎｄｏｗｓ（登録商標）系オペレーティングシステムが実行される情報処理装置２００によりＬＡＮ４００を介してアクセスされるデータをいう。追加されるとは、ＣＨＮ１（１１０）によりＮＦＳ／ＣＩＦＳファイル共有データがＬＵに新たに書き込まれることをいう。このとき当該ＮＦＳ／ＣＩＦＳファイル共有データに対応するメタデータもＬＵに書き込まれる。またＮＦＳファイルロック情報も更新される。

【０１１４】

次にＣＨＮ１（１１０）は、ＮＦＳファイルロック情報の同期処理を行う（S3

001）。図１６に示した同期方法管理テーブルの”同期方法欄”に記載されているように、ＮＦＳファイルロック情報は共有ＬＵに記憶される。そのためＣＨＮ１（１１０）は共有ＬＵに記憶されているＮＦＳファイルロック情報を更新する。なお、同期方法管理テーブルの”ローカルＬＵへの反映要否欄”に記載されているように、クラスタ内の他のＣＨＮ（１１０）のローカルＬＵへの反映は行われない。

【０１１５】

続いてS3002において、ＣＨＮ２（１１０）は上記更新されたＮＦＳファイルロック情報を確認する。必要に応じて自己のローカルＬＵに反映するようにすることもできる。

ＣＨＮ２（１１０）は、ＣＨＮ１（１１０）によるハートビートマークを確認し、一定時間を経過しているにも拘わらず未更新である場合や、ハートビートマークに障害発生を示す符号が記載されているものを発見した場合には、フェイルオーバ処理を開始する（S3003）。ハートビートマークは、ＣＨＮ１（１１０）及びＣＨＮ２（１１０）の双方が共有メモリ１２０に書き込むことにより、お互いの動作状態をチェックするためのデータである。

【０１１６】

S3004においては、ＣＨＮ２（１１０）は引き継ぎデータ参照テーブルの”データの格納先欄”を参照することにより、ＮＦＳファイルロック情報が共有ＬＵに記憶されていることを知ることができる。そして引き継ぎデータ参照テーブルの”データ通知有無欄”を参照し、通知がないことを知ることができる。通知がないのでＮＦＳファイルロック情報は共有ＬＵの所定の記憶位置に記憶されていることがわかり、ＣＨＮ２（１１０）は共有ＬＵの当該所定の記憶位置からＮＦＳファイルロック情報を読み出すことができる。このようにしてＣＨＮ２（１１０）はＣＨＮ１（１１０）からＮＦＳファイルロック情報を引き継ぐことができる。他の引き継ぎデータについても同様に、引き継ぎデータ参照テーブルを参照することによりＣＨＮ１（１１０）からＣＨＮ２（１１０）に引き継ぐことができる。これによりＣＨＮ２（１１０）は、それまでＣＨＮ１（１１０）により行われていた、情報処理装置２００に対するファイルアクセスサービスを引き継いで行

うことができるようになり、フェイルオーバが完了する（S3004）。

【０１１７】

このように、本実施の形態に係る記憶デバイス制御装置１００においては、引き継ぎデータの同期を取るようにすることにより、ＣＨＮ１１０に障害が発生してからの煩雑なデータ引き継ぎ処理を行う必要がなくなり、フェイルオーバを迅速に行うことが可能となる。またフェイルオーバの引き継ぎデータの種類に応じた特性を考慮して、最適な方法により同期処理を行うことができる。例えば、クラスタ内のＣＨＮ１１０でのみ同期が必要な引き継ぎデータについては共有ＬＵに書き込み、ストレージシステム６００全体のＣＨＮ１１０で同期が必要なデータについては全体共有ＬＵに書き込む。また共有ＬＵに書き込んだ引き継ぎデータを、他のＣＨＮ１１０のローカルＬＵにも書き込むようにすることができる。これにより、各ＣＨＮ１１０は引き継ぎデータを参照する際には自己のローカルＬＵにアクセスすればよく、共有ＬＵにアクセスする必要がなくなるので、他のＣＨＮ１１０とのアクセス競合の頻度を減少させることができ、データアクセス性能を向上させることができる。

【０１１８】

また本実施の形態における記憶デバイス制御装置１００においては、引き継ぎデータ参照テーブルを参照しながら引き継ぎデータの参照が行われることにより、引き継ぎデータの種類に応じてさまざまな位置に記憶されているフェイルオーバの引き継ぎデータを、適切に読み出すことが可能となる。

【０１１９】

以上本実施の形態について説明したが、上記実施例は本発明の理解を容易にするためのものであり、本発明を限定して解釈するためのものではない。本発明はその趣旨を逸脱することなく変更、改良され得ると共に、本発明にはその等価物も含まれる。

【０１２０】

【発明の効果】

記憶デバイス制御装置、及びプログラムを提供することができる。

【図面の簡単な説明】

【図１】　本実施の形態に係るストレージシステムの全体構成を示すブロック図である。

【図２】　本実施の形態に係る管理端末の構成を示すブロック図である。

【図３】　本実施の形態に係る物理ディスク管理テーブルを示す図である。

【図４】　本実施の形態に係るＬＵ管理テーブルを示す図である。

【図５】　本実施の形態に係るストレージシステムの外観構成を示す図である。

【図６】　本実施の形態に係る記憶デバイス制御装置の外観構成を示す図である。

【図７】　本実施の形態に係るチャネル制御部を示す図である。

【図８】　本実施の形態に係るメモリに記憶されるデータの内容を説明するための図である。

【図９】　本実施の形態に係るディスク制御部を示す図である。

【図１０】　本実施の形態に係るソフトウエア構成図である。

【図１１】　本実施の形態に係るチャネル制御部においてクラスタが構成されている様子を示す図である。

【図１２】　本実施の形態に係るメタデータを示す図である。

【図１３】　本実施の形態に係るロックテーブルを示す図である。

【図１４】　本実施の形態に係るストレージシステムにおけるシステムＬＵ、ユーザＬＵ、共有ＬＵを示す図である。

【図１５】　本実施の形態に係るストレージシステムにおけるローカルＬＵ、共有ＬＵ、全体共有ＬＵを示す図である。

【図１６】　本実施の形態に係るフェイルオーバ時に引き継がれるデータと同期の方法を示すテーブルである。

【図１７】　本実施の形態に係るフェイルオーバ時に引き継がれるデータの同期の方法を決定するための処理を示すフローチャートである。

【図１８】　本実施の形態に係るフェイルオーバ時に引き継がれるデータの参照先決定するためのテーブルである。

【図１９】　本実施の形態に係るフェイルオーバ時に引き継がれるデータの参

照先を決定するための処理を示すフローチャートである。

【図２０】　本実施の形態に係るフェイルオーバ処理を示すフローチャートである。

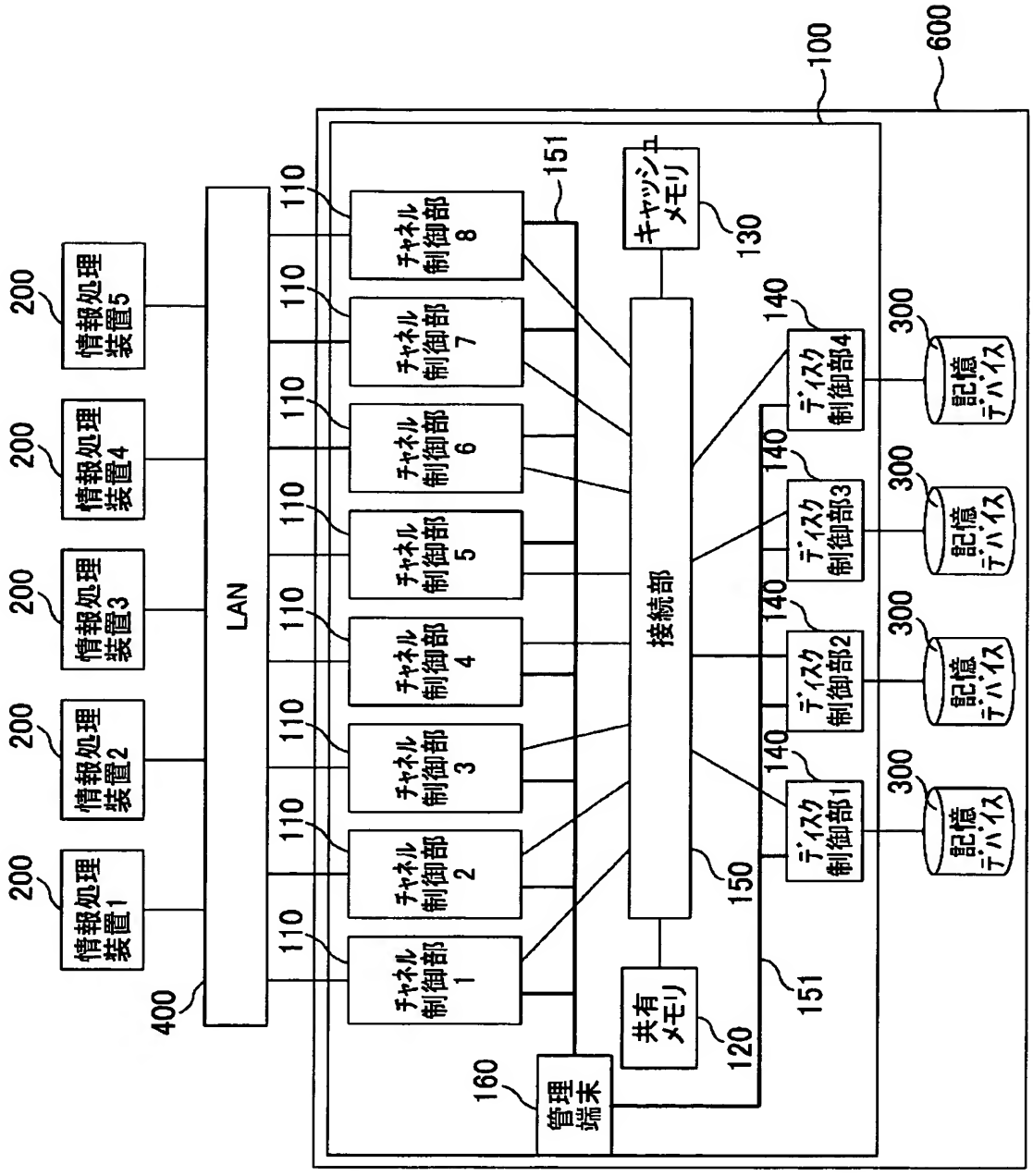
【符号の説明】

１００　記憶デバイス制御装置
１１０　チャネル制御部
１１１　ネットワークインタフェース部
１１２　ＣＰＵ
１１３　メモリ
１１４　入出力制御部
１１５　ＮＶＲＡＭ
１１６　ボード接続用コネクタ
１１７　通信コネクタ
１１８　回路基板
１１９　Ｉ／Ｏプロセッサ
１２０　共有メモリ
１３０　キャッシュメモリ
１４０　ディスク制御部
１５０　接続部
１５１　内部ＬＡＮ
１６０　管理端末
１７０　ファン
１８０　クラスタ
２００　情報処理装置
３００　記憶デバイス
４００　ＬＡＮ
６００　ストレージシステム
７３０　メタデータ
７２１　ファイルロックテーブル

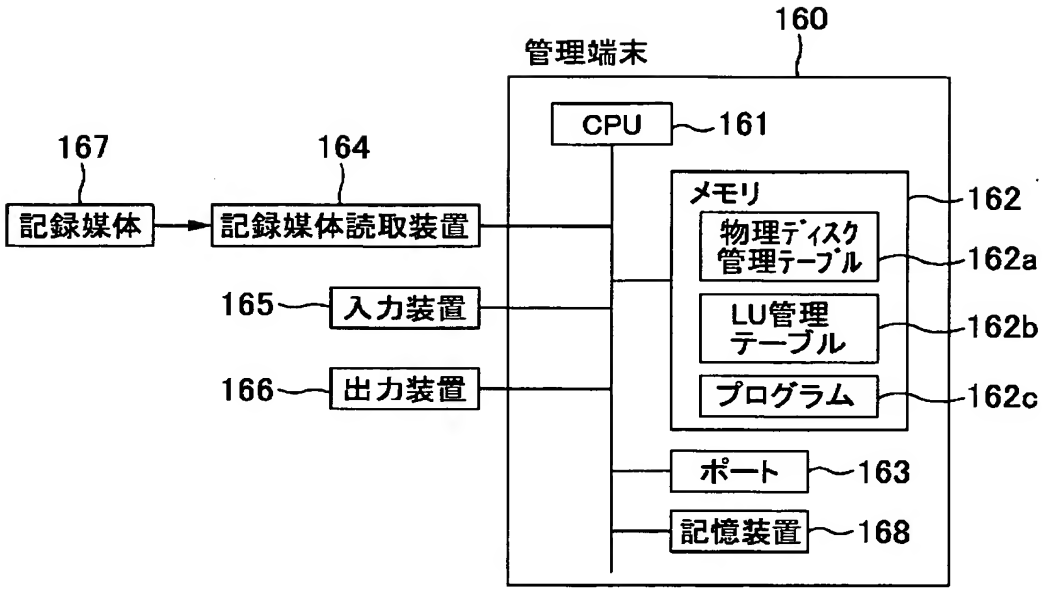
７２２　　ＬＵロックテーブル

【書類名】　図面

【図１】


100
600
LAN
400
情報処理装置1
情報処理装置2
情報処理装置3
情報処理装置4
情報処理装置5
200
チャネル制御部1
チャネル制御部2
チャネル制御部3
チャネル制御部4
チャネル制御部5
チャネル制御部6
チャネル制御部7
チャネル制御部8
110
151
キャッシュメモリ
130
接続部
150
共有メモリ
120
管理端末
160
ディスク制御部1
ディスク制御部2
ディスク制御部3
ディスク制御部4
140
記憶デバイス
300

【図２】



【図３】

162a

物理ディスク管理テーブル

ディスク番号	容量	RAID	使用状況
#001	100GB	5	使用中
#002	100GB	5	使用中
#003	100GB	5	使用中
#004	100GB	5	使用中
#005	100GB	5	使用中
#006	50GB	―	未使用
⋮	⋮	⋮	⋮

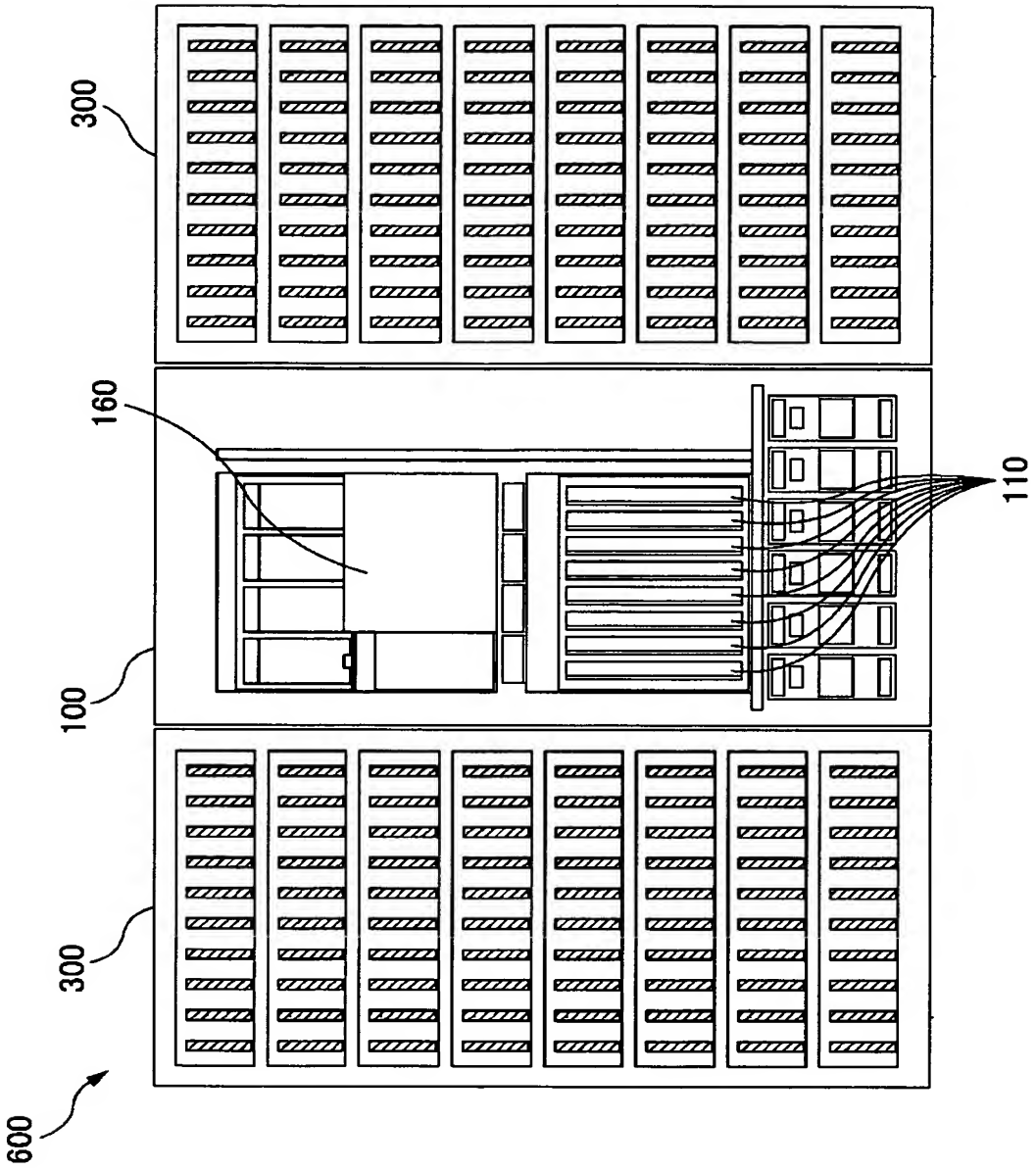
【図４】

162b

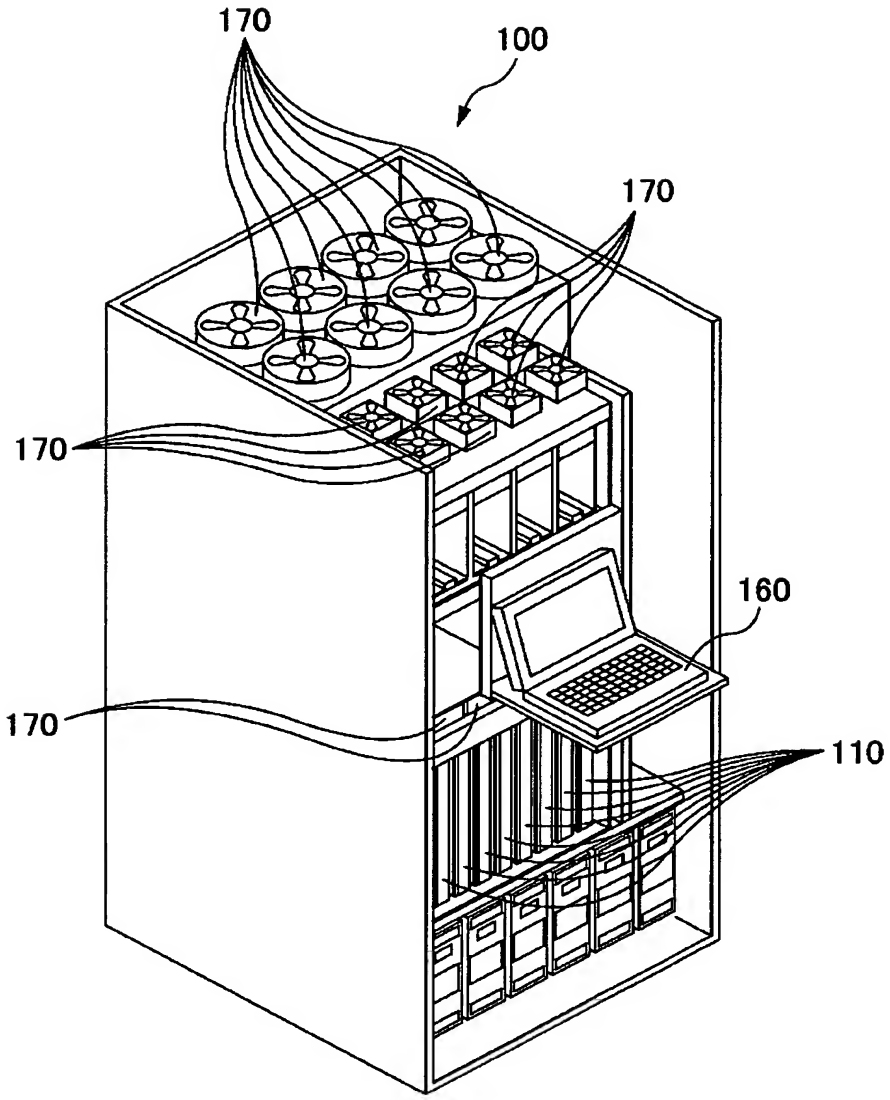
LU管理テーブル

LU番号	物理ディスク	容量	RAID
#1	#001,#002,#003,#004,#005	100GB	5
#2	#001,#002,#003,#004,#005	300GB	5
#3	#006,#007,	200GB	1
⋮	⋮	⋮	⋮

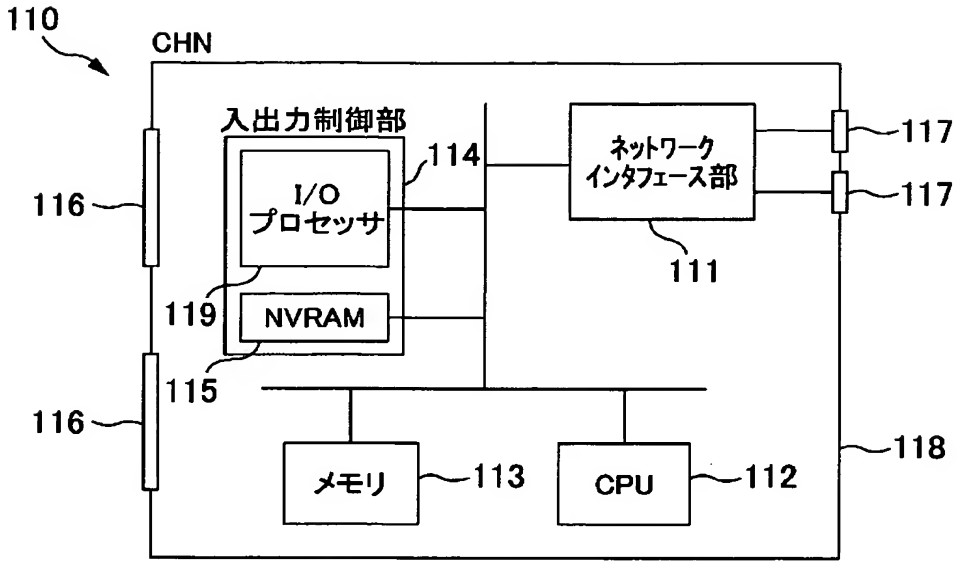
【図５】



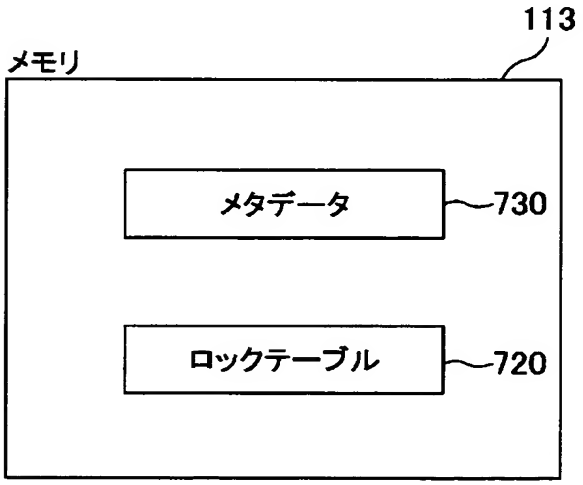
【図６】



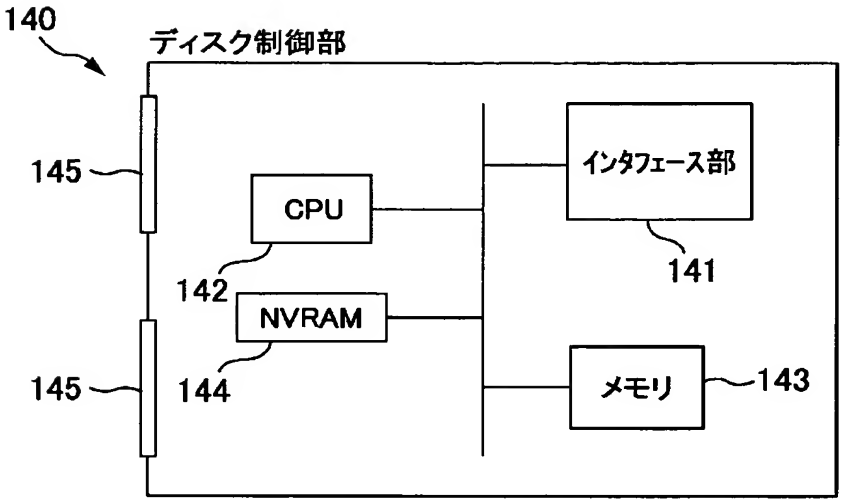
【図７】


110
CHN
入出力制御部
I/O
プロセッサ
114
NVRAM
119
115
116
116
ネットワーク
インタフェース部
111
117
117
メモリ
113
CPU
112
118

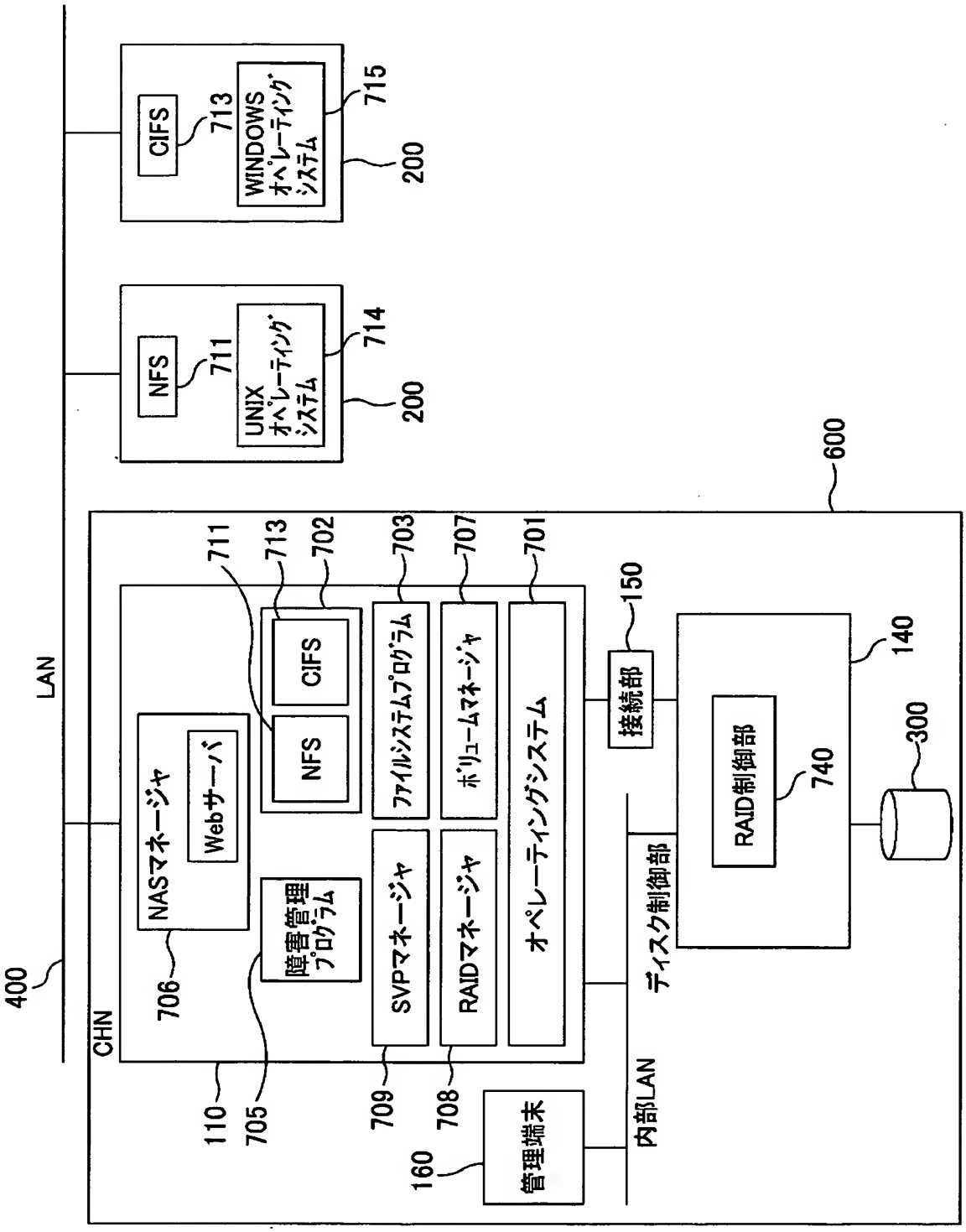
【図８】


113
メモリ
メタデータ
730
ロックテーブル
720

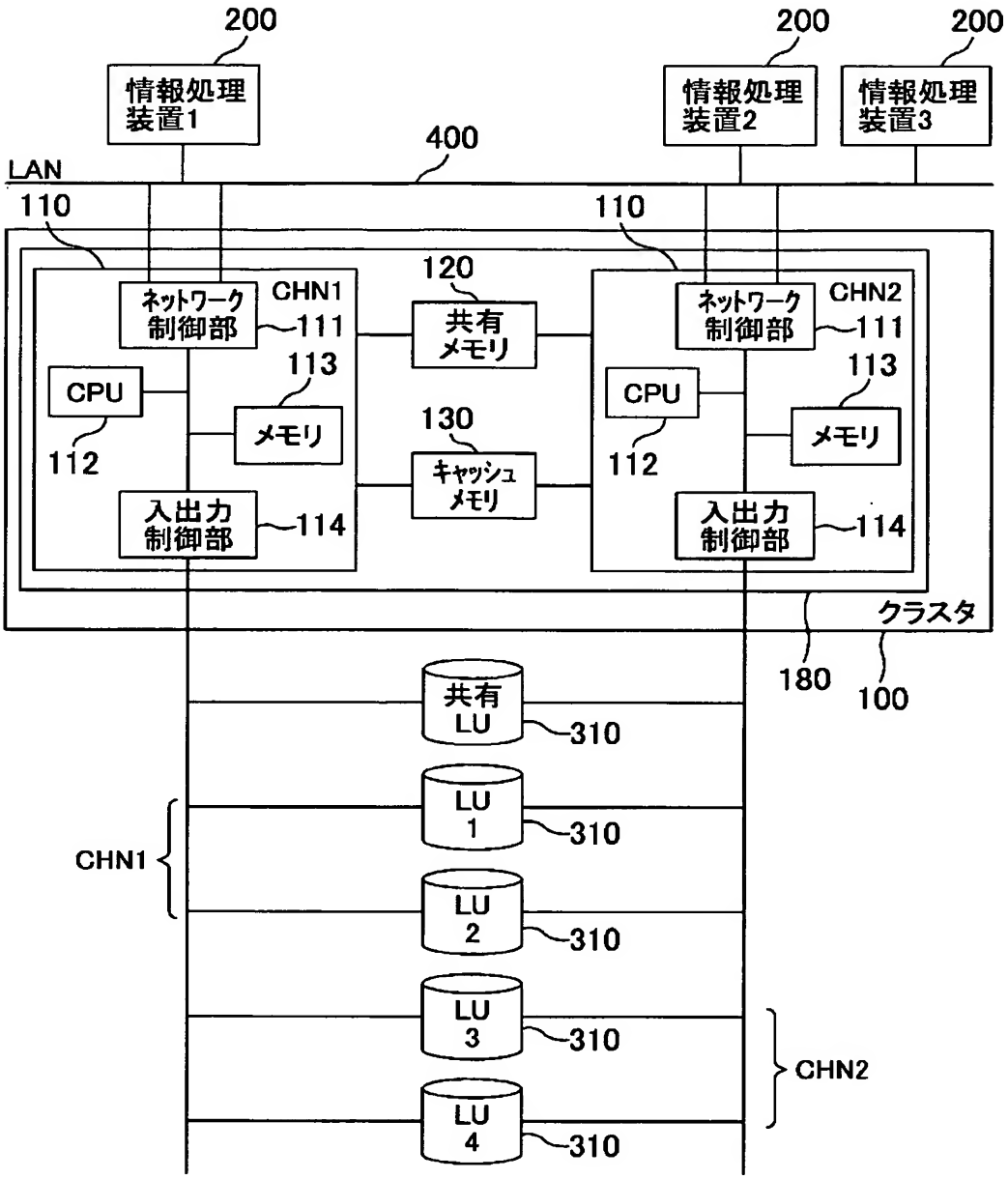
【図９】



【図１０】


CIFS
713
WINDOWS
オペレーティング
システム
715
200
NFS
711
UNIX
オペレーティング
システム
714
200
600
LAN
711
713
702
703
707
701
150
140
CIFS
NFS
ファイルシステムプログラム
ボリュームマネージャ
オペレーティングシステム
接続部
RAID制御部
740
300
400
NASマネージャ
Webサーバ
706
障害管理
プログラム
SVPマネージャ
RAIDマネージャ
ディスク制御部
CHN
110
705
709
708
160
管理端末
内部LAN

【図１１】


200
情報処理装置1
200
情報処理装置2
200
情報処理装置3
400
LAN
110
110
ネットワーク制御部
CHN1
111
120
共有メモリ
ネットワーク制御部
CHN2
111
113
CPU
メモリ
130
CPU
113
メモリ
112
キャッシュメモリ
112
入出力制御部
114
入出力制御部
114
クラスタ
180
100
共有LU
310
LU 1
310
CHN1
LU 2
310
LU 3
310
CHN2
LU 4
310

【図１２】

730

メタデータ

ファイル名	先頭アドレス	容量	所有者	更新時刻
A	7BSA	200MB	X	0:00
B	05BF	50MB	X	7:57
C	1F30	100MB	Y	9:15
D	470B	100MB	Z	15:20
⋮	⋮	⋮	⋮	⋮

【図１３】

721

ファイルロックテーブル

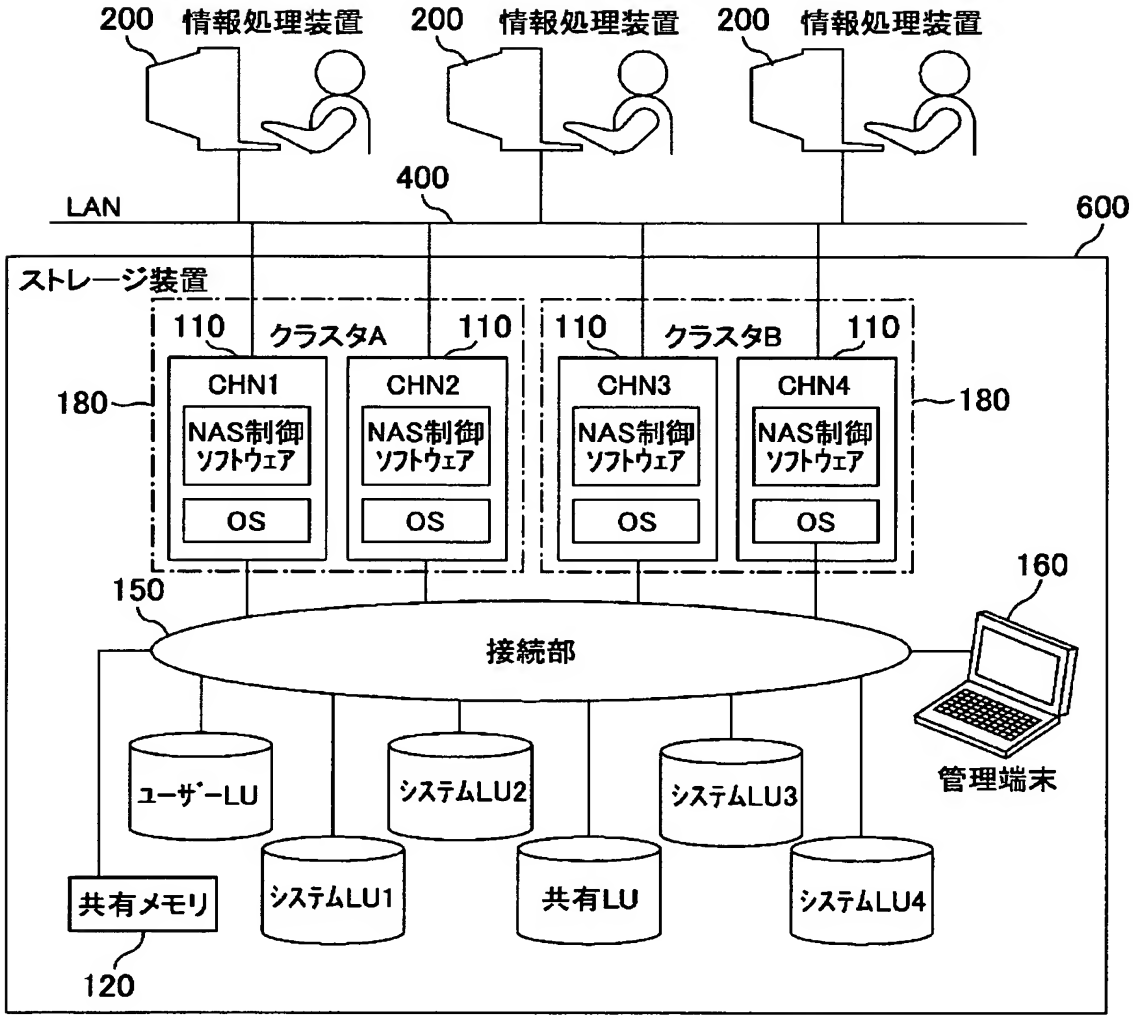
ファイル名	ロック状態
A	ロック中
B	—
C	—
D	ロック中
⋮	⋮

722

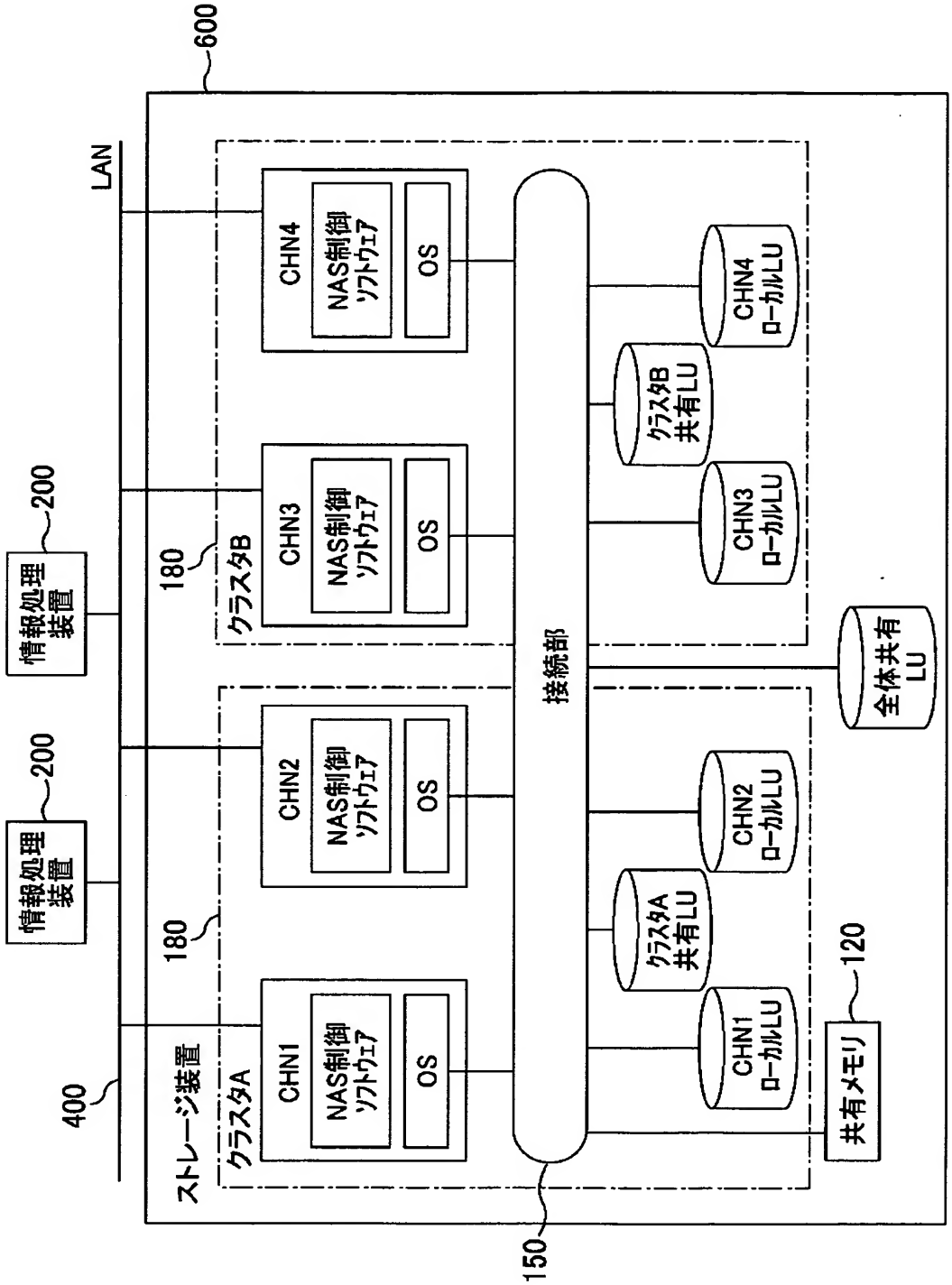
LUロックテーブル

LU	ロック状態
共有	—
1	ロック中
2	—
⋮	⋮

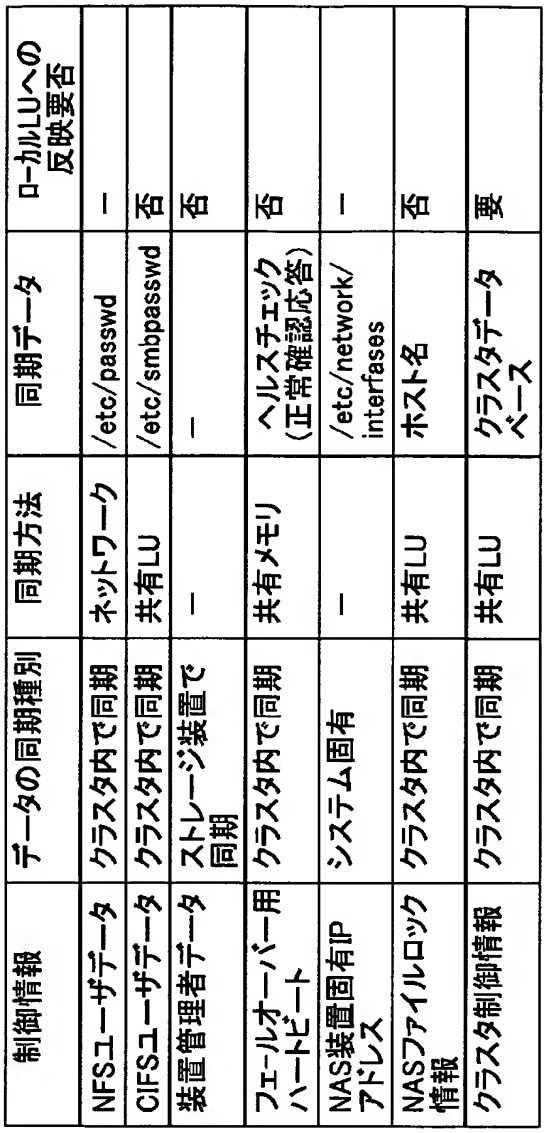
【図１４】


200
情報処理装置
200
情報処理装置
200
情報処理装置
400
LAN
600
ストレージ装置
110
クラスタA
110
110
クラスタB
110
CHN1
CHN2
CHN3
CHN4
180
180
NAS制御ソフトウェア
NAS制御ソフトウェア
NAS制御ソフトウェア
NAS制御ソフトウェア
OS
OS
OS
OS
150
160
接続部
ユーザーLU
システムLU2
システムLU3
管理端末
共有メモリ
システムLU1
共有LU
システムLU4
120

【図１５】

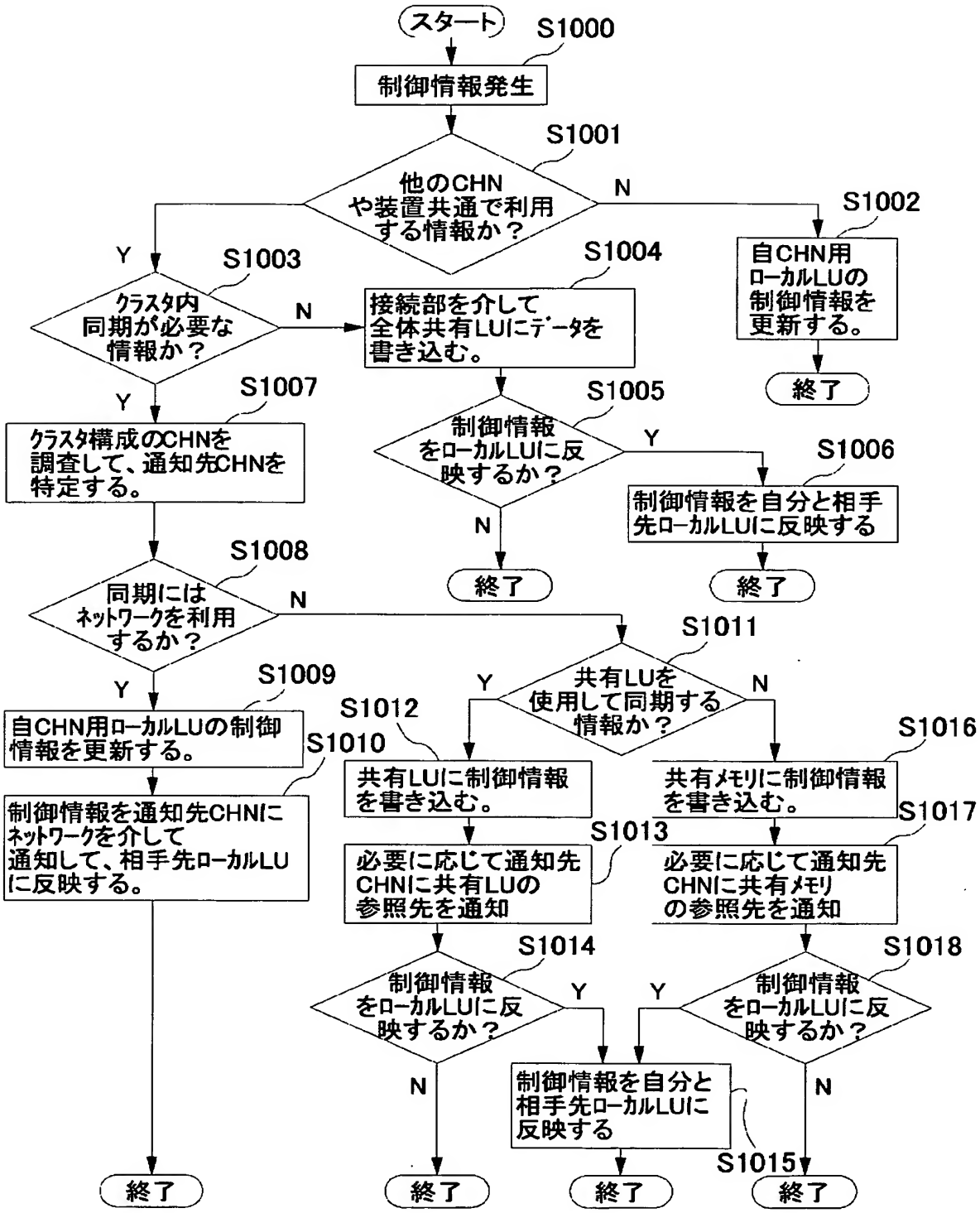

600
LAN
情報処理装置
200
情報処理装置
200
400
ストレージ装置
クラスタA
クラスタB
180
180
CHN1
NAS制御ソフトウェア
OS
CHN2
NAS制御ソフトウェア
OS
CHN3
NAS制御ソフトウェア
OS
CHN4
NAS制御ソフトウェア
OS
接続部
150
CHN1ローカルLU
CHN2ローカルLU
CHN3ローカルLU
CHN4ローカルLU
クラスタA共有LU
クラスタB共有LU
全体共有LU
共有メモリ
120

【図１６】



制御情報	データの同期種別	同期方法	同期データ	ローカルLUへの反映要否
NFSユーザデータ	クラスタ内で同期	ネットワーク	/etc/passwd	—
CIFSユーザデータ	クラスタ内で同期	共有LU	/etc/smbpasswd	否
装置管理者データ	ストレージ装置で同期	—	—	否
フェールオーバー用ハートビート	クラスタ内で同期	共有メモリ	ヘルスチェック（正常確認応答）	否
NAS装置固有IPアドレス	システム固有	—	/etc/network/interfases	—
NASファイルロック情報	クラスタ内で同期	共有LU	ホスト名	否
クラスタ制御情報	クラスタ内で同期	共有LU	クラスタデータベース	要

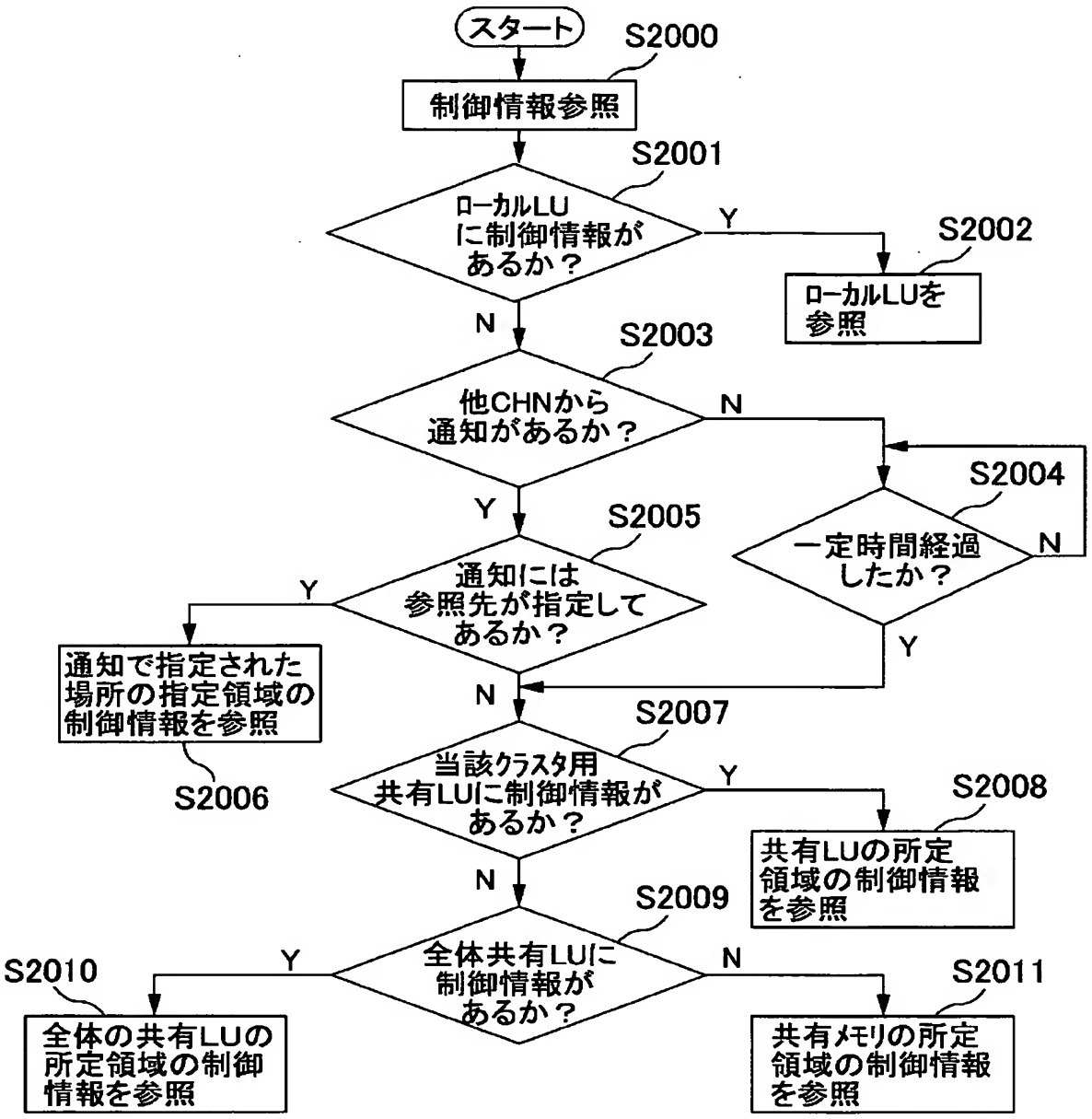
【図１７】


スタート
S1000
制御情報発生
S1001
他のCHN
や装置共通で利用
する情報か？
N
S1002
自CHN用
ローカルLUの
制御情報を
更新する。
終了
Y
S1003
クラスタ内
同期が必要な
情報か？
N
S1004
接続部を介して
全体共有LUにデータを
書き込む。
S1005
制御情報
をローカルLUに反
映するか？
Y
S1006
制御情報を自分と相手
先ローカルLUに反映する
終了
N
終了
Y
S1007
クラスタ構成のCHNを
調査して、通知先CHNを
特定する。
S1008
同期には
ネットワークを利用
するか？
N
S1011
共有LUを
使用して同期する
情報か？
Y
N
Y
S1009
自CHN用ローカルLUの制御
情報を更新する。
S1010
制御情報を通知先CHNに
ネットワークを介して
通知して、相手先ローカルLU
に反映する。
S1012
共有LUに制御情報
を書き込む。
S1016
共有メモリに制御情報
を書き込む。
S1013
必要に応じて通知先
CHNに共有LUの
参照先を通知
S1017
必要に応じて通知先
CHNに共有メモリ
の参照先を通知
S1014
制御情報
をローカルLUに反
映するか？
Y
S1018
制御情報
をローカルLUに反
映するか？
Y
N
制御情報を自分と
相手先ローカルLUに
反映する
S1015
N
終了
終了
終了
終了

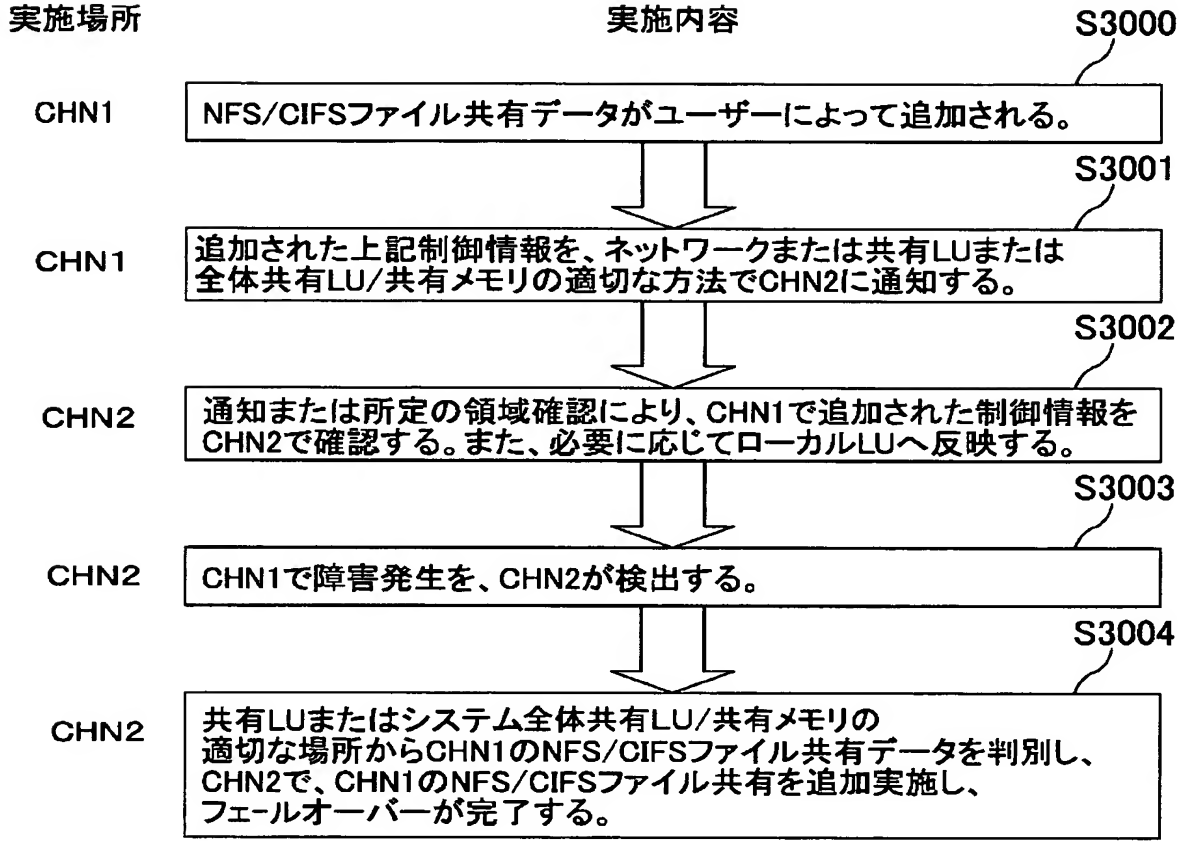
【図１８】

制御情報	データの格納先	データの通知有無
NFSユーザデータ	ローカルLU	ー
CIFSユーザデータ	共有LU	あり
装置管理者データ	全体共有LU	なし
フェールオーバー用ハートビート	共有メモリ	なし
NFSファイルロック情報	共有LU	なし
クラスタ制御情報	ローカルLU	なし

【図１９】


スタート
S2000
制御情報参照
S2001
ローカルLUに制御情報があるか？
Y
S2002
ローカルLUを参照
N
S2003
他CHNから通知があるか？
N
S2004
一定時間経過したか？
N
Y
Y
S2005
通知には参照先が指定してあるか？
Y
通知で指定された場所の指定領域の制御情報を参照
S2006
N
S2007
当該クラスタ用共有LUに制御情報があるか？
Y
S2008
共有LUの所定領域の制御情報を参照
N
S2009
全体共有LUに制御情報があるか？
Y
S2010
全体の共有LUの所定領域の制御情報を参照
N
S2011
共有メモリの所定領域の制御情報を参照

【図２０】


実施場所
実施内容
S3000
CHN1
NFS/CIFSファイル共有データがユーザーによって追加される。
S3001
CHN1
追加された上記制御情報を、ネットワークまたは共有LUまたは
全体共有LU/共有メモリの適切な方法でCHN2に通知する。
S3002
CHN2
通知または所定の領域確認により、CHN1で追加された制御情報を
CHN2で確認する。また、必要に応じてローカルLUへ反映する。
S3003
CHN2
CHN1で障害発生を、CHN2が検出する。
S3004
CHN2
共有LUまたはシステム全体共有LU/共有メモリの
適切な場所からCHN1のNFS/CIFSファイル共有データを判別し、
CHN2で、CHN1のNFS/CIFSファイル共有を追加実施し、
フェールオーバーが完了する。

【書類名】　要約書

【要約】

【解決手段】　情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対するデータ入出力要求に対応するＩ／Ｏ要求を出力するＩ／Ｏプロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、各チャネル制御部はフェイルオーバを行うためのグループに類別されている記憶デバイス制御装置であって、各チャネル制御部により更新されるフェイルオーバ時に引き継がれるデータを、当該チャネル制御部と同一のグループに類別されている各チャネル制御部が共通にアクセス可能な記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である共有ボリュームに記憶する手段を備える。

【選択図】　図１

特願２００３－０２５０７４

出願人履歴情報

識別番号　［０００００５１０８］

１．変更年月日　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所